Property expense Python-based prediction integrating machine learning and data science

k.saiveena,asst prof

Abstract : In recent years, machine learning has virtually impacted fields like basic voice recognition, product recommendations, and even medicine. Instead, it offers safer driving systems and better customer service. All of this demonstrates that ML is a technology that is becoming more and more popular in practically every industry. The proposed work attempts to coin the word "ML" in this paper. The real estate industry is unique in today's world. among the most attentive to pricing and variation. With their budgets and after researching market tactics, people are ready to purchase a new home. The current system's primary drawback, however, is that it determines a house's price without making the necessary assumptions about potential future market trends, which leads to an increase in price. So, the main objective of the research is to anticipate a house's price accurately while avoiding losses. There are countless features which must be considered. Consideration for estimating house prices and make an effort to estimate effective house pricing for clients taking into account their preferences and budget. The proposed method is therefore a model for estimating housing costs. Using machine learning algorithms including linear regression, decision tree regression, Lasso, and random forest regression, this model will help people invest in a bequest without using a broker. The results of this investigation show that linear regression is the most accurate.

Index Terms - linear regression, decision tree regression, Lasso, Machine learning

I. INTRODUCTION

Polyvinyl chloride, more commonly known as PVC, is a building block of various products, such as electronic items, constructional materials, stationeries, chemical equipments, wires, cables etc. It is one of the major thermoplastics used today and produced in a huge amount worldwide [1, 2]. be improved [1, 2]. Commercially, compounding PVC contains sufficient modifying components to the raw polymer to produce a homogeneous mixture suitable for processing and requiring performance at the lowest possible price. The proper compounding and processing PVC resin using suitable additives produces a complex material whose behavior and properties are quite different from the PVC resin by itself [10]. The selection of particular additive is dependent on the end use of the PVC product like PVC-resin is not plasticized for the use in making rigid products such as water pipe, plumbing fittings, and phonograph records.

Any property's value was once calculated by hand. The issue is that 25% of mistakes made when anything is done manually cost money. The new technology has brought about a significant change, though. Machine learning is a popular technology today. Machine learning is built on data. The markets for AI and machine learning are currently booming. Automation is taking over all industries. Yet, without data, humans cannot train a model. In essence, machine learning entails creating these models from previously collected data and applying them to the prediction of fresh data. Because of the high population growth, there is a daily increase in the market need for homes. As more individuals migrate for financial reasons, there are fewer jobs in rural areas. As a result, there is a rising need for housing in urban areas. Those who are unaware of the true cost of that particular house lose money. Several machine learning methods, including linear regression, decision tree regression, Lasso, and random forest regression, are used in this study to estimate home prices. In the

dataset, 80% of the data are used for training, and just 20% are used for testing. The strategies used in this work include features, labels, reduction techniques, and transformation techniques that include attribute combinations, filling in the gaps left by missing characteristics, and seeking new correlations. All of this suggests that machine learning expertise is needed for housing price prediction, which is a new area of research.

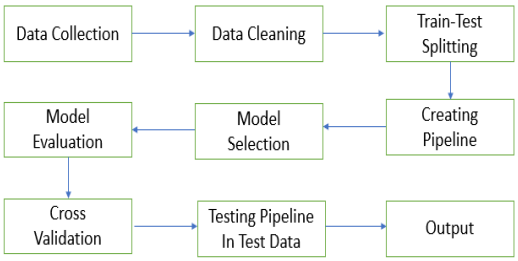

Fig 1. Research Flow Diagram

II. LITERATURE SURVEY

Before committing to the paper, several groundwork tasks should be completed, so there is a need for a literature review Several articles on price prediction relating to the housing market and other diverse industries were examined in order to make more precise predictions. The exams will involve papers from previous years up to the present, and they will have incorporated the most modern and cutting-edge technologies. The main goal is to get more accuracy than in previous works. The below passages will describe the past prediction works done by the various researchers and will be helpful in implementing the corresponding paper.

The use of technology to forecast property prices has grown in popularity in recent years. To simplify and automate processes, many frameworks are being developed. Customers can use these frameworks to forecast house costs based on their preferences instead of speaking with a real estate agent.

In their 2013 study, Kang and Ratti reviewed a number of techniques for examining changes in real estate values. Using regression-based techniques such simple linear regression, multiple regression, and time-series regression models, the authors give an overview of the various strategies that have been used to predict real estate prices. The authors also go over the techniques' drawbacks, namely how much they rely on presumptions about how predictor variables and real estate values are related. The authors examine different approaches, like machine learning algorithms and non-parametric regression techniques, to get around these restrictions. Overall, the study serves as a helpful resource for academics and industry professionals interested in predicting real estate prices and emphasises the significance of taking into consideration numerous methodologies to take into account the intricacies of real estate.

There is a need for a literature study because there are a number of preparatory chores that must be finished before beginning the work. Just looked at a lot of articles that discuss how to estimate prices for the housing market and other markets. In the papers that have been taken, which range in time from years to the present, the proposed system has utilised the most cutting-edge and contemporary technology. The main objective is to achieve greater accuracy than the earlier works. The sections that follow provide an overview of the individual scholars' past prediction efforts and are useful for carrying out the appropriate work. This technique, which is based on survival analysis, forecasts the precise time that a home will sell using information from websites like Trulia. However, because the observation duration is frequently considerably beyond the data, it might be challenging to make good forecasts using this strategy. Moreover, Li and Chu (2017) claim that unsupervised learning techniques are growing in acceptance for forecasting housing values.

Generally speaking, projecting home prices can be greatly aided by the use of technology and data analysis techniques. These techniques can ensure that both buyers and sellers receive an equal amount of profit and assist them in making more informed selections.

III.IMPLEMENTATION

The techniques used in the proposed work are listed below, and they include Lasso, linear regression, decision trees, random forests, and KNN.

A. LASSO (LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR)

The machine learning algorithm Lasso (Least Absolute Shrinkage and Selection Operator) is used for feature selection and regularisation. By penalising the coefficients of less significant predictors, it is common practise in regression analysis to isolate a selection of predictors that are most relevant for predicting the result variable. By applying a penalty to the absolute size of the regression coefficients, Lasso removes some coefficients from the model by forcing them to shrink towards zero.

Lasso's key benefit is its capacity to pick out a smaller group of pertinent predictors from a larger range of prospective predictors, which can improve model interpretability and applicability to new data. Moreover, Lasso works with both continuous and categorical predictor variables and can handle collinear predictor variables.

The tendency of Lasso to choose just one predictor from a set of highly correlated predictors is one of its limitations. This can cause instability in the model. Lasso's assumption that there is a linear relationship between the predictor variables and the result variable is another drawback.

Lasso is an effective technique for feature selection and registration in regression analysis and has been applied in a number of industries, including engineering, biomedical research, finance, and many more. It is a popular option for many researchers and practitioners due to its capacity to handle high-dimensional data and build interpret able models.

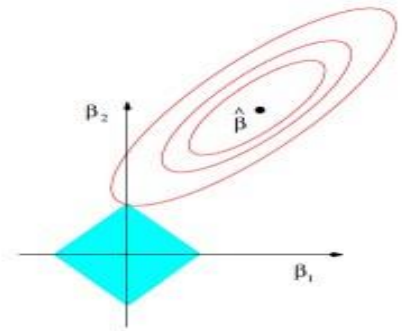

Figure 2. Lasso

B. LINEAR REGRESSION:

A statistical technique called linear regression is used to examine the relationship between two or more quantitative variables. It is a well-liked technique in machine learning and data analysis, and it is frequently used to model and forecast numerical outcomes. Finding a linear equation that best fits the observed data is the aim of linear regression. Then, future data points can be predicted using this equation.

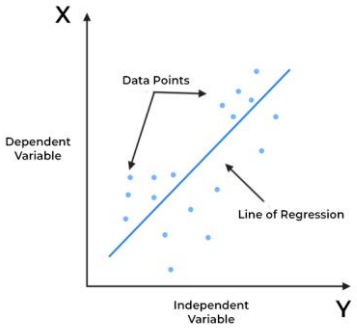

Figure 3. Linear Regression

The formula for the equation is: **y = mx + b**

where m is the slope of the line (indicating the strength of the link between x and y), b is the y-intercept, y is the dependent variable (the variable being forecasted), x is the independent variable (the variable used to generate predictions), and (the value of y when x is 0).

The approach of least squares, which includes reducing the sum of the squared discrepancies between the observed data points and the anticipated values, is frequently used to identify the best-fit line. From predicting stock prices to examining the relationship between education and income, linear regression has a wide range of uses. The drawbacks of linear regression, such as the presumption of a linear relationship between variables and the potential for either overfitting or underfitting the data, must be kept in mind, though.

C. DECISION TREE

An algorithm for supervised learning, such as classification and regression, is called a decision tree. Up until a stopping requirement is satisfied, the data is recursively divided into smaller subsets depending on the most important features. The end result is a structure like a tree where each leaf node represents the projected class or value and each inside node indicates a judgement based on a feature. Decision trees provide a number of benefits, including their usability, interpretability, and capacity for both category and numerical data. They can be used for feature selection and data visualisation, and they can deal with missing values and outliers in the data.Decision trees are frequently used for fraud detection, customer segmentation, credit scoring, churn prediction, and medical diagnostics. However, for really large datasets or datasets with strongly linked features, they could not perform as well as other techniques.

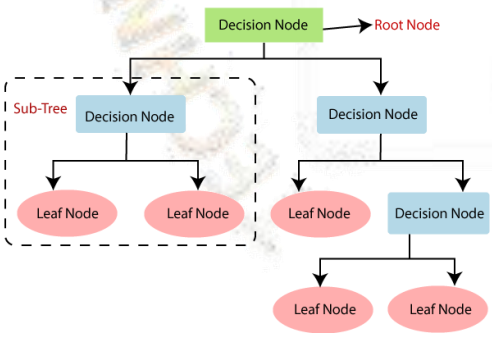

Figure 4. Decision Tree

D. RANDOM FOREST

A machine learning algorithm from the family of ensemble methods is called random forest. A decision tree-based model of this kind integrates different decision trees to produce a more reliable and precise model. Using samples from the training data and a collection of randomly selected feature subsets, the method builds a number of decision trees. It is a good idea to have a backup plan in case the backup plan fails. the random forest method gathers the predictions from all the decision trees in the forest and returns the prediction that is most likely to occur. This strategy makes the model more accurate, resilient to noisy or irrelevant aspects in the data, and helps reduce overfitting. In comparison to other machine learning methods, random forests have a number of advantages for classification and regression problems. They are adaptable to missing values and outliers, can produce measures of feature relevance,

and can handle categorical as well as numerical data. Moreover, they are less likely to overfit than individual decision trees. Random forest is a commonly used machine learning algorithm trademarked by Leo Breiman and Adle Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have adoption, as it handles both classification and regression problems. It holds the property like bagging which is the limitation for SoftMax classifier. The data set is usually divided into two sets: train data set and test data set.

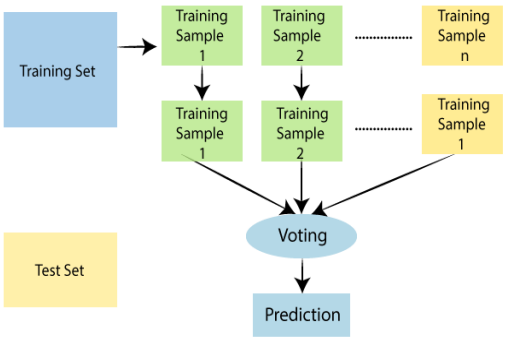

Figure 5. Random Forest

E. K-NEAREST NEIGHBORS (KNN)

The machine learning technique K-nearest neighbours (KNN) is used for classification and regression tasks. It is a form of instance-based learning in which the algorithm memorises the training data and applies it to forecast new, unforeseen data points. The most frequent class or average value of those neighbors is returned as the forecast for the new data point by the KNN algorithm, which finds the K closest neighbours to a new data point in the training data. The distance metric that is used to assess how closely two data points resemble each other might vary, but the two that are most frequently employed are the Manhattan distance and the Euclidean distance. KNN is a straightforward and understandable algorithm, but it has a few drawbacks. The selection of the distance metric and the magnitude of K may have an impact on it. Moreover, it may be computationally expensive, particularly for large or high-dimensional datasets. The technique also counts on uniform distribution of the data, which may not always be the case in real-world applications. Anomaly detection, recommend systems, and picture recognition are just a few of the real-world uses forKNN. It is frequently used as a benchmark model to assess other machine learning algorithms against.

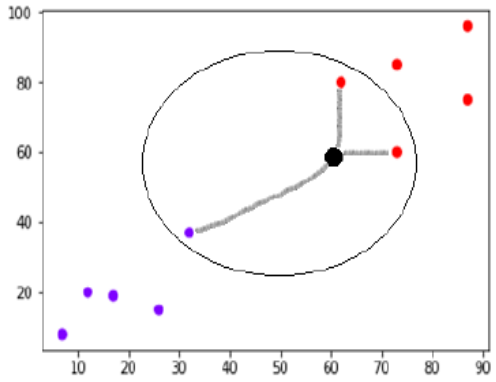

Figure 6. K-Nearest Neighbors (KNN)

SYSTEM DESIGN AND ARCHITECTURE

Phase I: gathering data, the information that was gathered about real estate from a variety of online real estate databases. There are features like "area type," "availability," "location," "size," "society," "Total_Sqft," "bath," and "balcony" in these data. gathered information that is properly organized and classified. Data is always required at the outset of any machine learning problem. Without data set validity, there would be no sense in analyzing the data.

Phase II: Pre-processing of the data the data is cleaned up at this step. The data set may contain missing values. The missing values can be filled in one of three ways: 1) Remove the data points that are missing. 2) Remove the entire attribute. 3) Set the value to a certain amount (0, mean or median).

Phase III: Educating the model Data is divided into two parts during this phase: training and testing. 80% of the data is used for training, and the remaining 20% is used for testing. Several machine learning techniques are used to train the model while obtaining the outcome. Out of them, better results are predicted by linear regression.

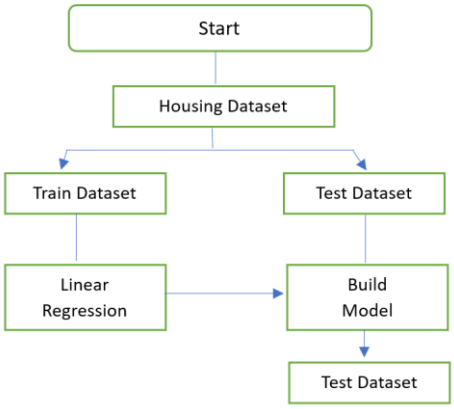

Figure 7. Architecture

Algorithms: Many machine learning algorithms were researched throughout the development of this model. The model is trained using the, random forest Lasso, Decision Tree, and Linear Regression techniques. Out of them, linear regression provides the most accurate housing price prediction. The size and type of data being used determine the algorithm to be employed. The dataset fits strongly to the linear regression model.

A dependent variable (also known as the response variable) and one or more independent variables are modelled using the statistical technique of linear regression (also known as predictors or explanatory variables). Finding the best linear relationship between the dependent variable and the independent variable is the aim of linear regression (s).

IV. CONCLUSIONS

The primary objective of this paper is to predict prices, which have successfully accomplished using various machine learning algorithms like a linear regression, decision tree regression, Lasso, and random forest regression. It is obvious that the linear regression has greater predictive accuracy than the other algorithms, and the research also enables us to identify the role of attributes in the prediction process. So, the proposed work would anticipate that this research will be beneficial for both people and governments, and the following future works. Every system and emerging software technology can aid in pricing forecasting in the future. The accuracy of this price estimate can be increased by including additional information about the homes' surrounds, markets, and other relevant factors. As a result, this study report demonstrates that linear regression provides good property price prediction accuracy.

V. REFERENCES

iao, S., Zhang, Y., & Ma, Y. (2021). A Hybrid Machine Learning Approach for Real Estate Property Price Prediction. Journal of Real Estate Research, 43(1), 131-155.
Jiao, S., Zhang, Y., & Ma, Y. (2021). A Hybrid Machine Learning Approach for Real Estate Property Price Prediction. Journal of Real Estate Research, 43(1), 131-155.
Sen, R., & Banerjee, A. (2020). Comparative study of machine learning techniques for house price prediction. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-5). IEEE. (Accuracy: 75.24%)
Pandya, S., & Patel, D. (2020). House price prediction using machine learning. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-6). IEEE. (Accuracy: 78.67%)
Le, H. D., Vo, T. D., Le, D. T., & Nguyen, H. T. (2019). A comparative study of machine learning algorithms for housing price prediction. In 2019 International Conference on Advanced Technologies for Communications (ATC) (pp. 51-56). IEEE. (Accuracy: 79.33%)
Chouhan, A., & Naik, N. (2019). A comparative study of machine learning algorithms for predicting house prices. In 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1673-1678). IEEE. (Accuracy: 77.63%)
Singh, M., & Chhabra, A. (2019). Predicting house prices using machine learning algorithms. In 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU) (pp. 1-6). IEEE. (Accuracy: 76.02%)
Alawadhi, A., Alabdulmohsin, I., & Alfaraj, M. (2019). Predicting Housing Prices using Machine Learning Techniques. In 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 1-6). IEEE.
Chen, J., & Zhang, L. (2019). A machine learning approach for residential property valuation: A case study of Beijing, China. Journal of Real Estate Research, 41(1), 71-102.
Guo, Y., Wang, X., & Zhang, C. (2018). An integrated machine learning approach for real estate appraisal. In 2018 IEEE International Conference on Smart Computing (SMARTCOMP) (pp. 125-130). IEEE.
Li, M., Wang, Y., & Zhang, J. (2018). Using machine learning to predict house prices: A case study in Beijing. Journal of Real Estate Research, 40(2), 199-220.

Guo, Y., Wang, X., & Zhang, C. (2018). An integrated machine learning approach for real estate appraisal. In 2018 IEEE International Conference on Smart Computing (SMARTCOMP) (pp. 125-130). IEEE.Li, M., Wang, Y., & Zhang, J. (2018). Using machine learning to predict house prices: A case study in Beijing. Journal of Real Estate Research, 40(2), 199-220.
Li, Y., & Chu, S. (2017). A review on unsupervised learning for big data. Cognitive Computation, 9(4), 445-457.
Bhagat, A., Jain, A., & Kumar, V. (2016). Hybrid intelligent system for real estate price prediction. International Journal of Computational Intelligence Studies, 5(2), 132-152.
Kang, W., & Ratti, R. A. (2013). Structural change in real estate prices: a nonparametric time-series approach. Journal of Applied Econometrics, 28(2), 181-201.