“House Price Prediction Using Machine Learning”

A

Project Report

submitted in partial fulfillment of the

requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

by

Roll No.	SAP ID	Name
R103219030	500077712	Shreya Sharma
R103219018	500075219	Shivani Saran
R134219146	500076504	Tanisha Pathak

under the guidance of

Sir Bikram Pratim Bhuyan

UNIVERSITY WITH A PURPOSE

Department of Informatics

School of Computer Science

University of Petroleum & Energy Studies

Bidholi, Via Prem Nagar, Dehradun, UK

May- 2022

CANDIDATE'S DECLARATION

We hereby certify that the project work entitled "**House Price Prediction Using Machine Learning**" in partial fulfillment of the requirements for the award of the Degree of **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE AND ENGINEERING** with specialization in **Business Analytics & Optimization** and submitted to the **Department of Informatics** at School of Computer Science, University of Petroleum & Energy Studies, Dehradun, is an authentic record of our work carried out during a period from August,22 to December,22 under the supervision of **Mr. Bikram Pratim Bhuyan**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: 12th Dec 2022

Bikram Pratim Bhuyan
School of Computer Science
University of Petroleum & Energy Studies
Dehradun – 248 001 (Uttarakhand)

ACKNOWLEDGEMENT

We wish to express our deep gratitude to our guide **Bikram Pratim Bhuyan**, for all advice, encouragement and constant support he has given us throughout our project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thank our respected Head of the Department, **Dr.Thipendra Pal Singh**, for his great support in doing our major project in **Area (like data cleansing, OOP etc.)** at **SoCS**.

We are also grateful to **Dr. Ravi Shankar Iyer**, **Dean SoCS**, for giving us the necessary facilities to carry out our project work successfully.

We would like to thank all our friends for their help and constructive criticism during our project work. Finally we have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

Name	**Shreya Sharma**	**Shivani Saran**	**Tanisha Pathak**
Roll No.	**R103219030**	**R103219018**	**R134219146**

ABSTRACT

Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of our project. We are doing how to predict housing costs using various regression techniques using the Python library. The proposed method takes into account the sophisticated aspects used in the house price calculation and provides a more accurate forecast. we uses machine learning to explain how the house price model works and which datasets are used in the proposed model. Predictive models to determine the selling prices of homes in cities like Bangalore are maintained because of more than difficult and sophisticated tasks. The price of real estate for sale in cities depends on several factors involved.

Data mining is now commonly applied in the real estate market. Data mining's ability to extract relevant knowledge from raw data makes it very useful to predict house prices, key housing attributes, and many more. Research has stated that the fluctuations in house prices are often a concern for house owners and the real estate market. A survey of literature is carried out to analyze the relevant attributes and the most efficient models to forecast the house prices.

The findings of this analysis verified the use of the linear regression, Bayesian Ridge as the most efficient models compared to Ridge, Lasso, decision tree, Random forest, Elastic net. Moreover, our findings also suggest that locational attributes and structural attributes are prominent factors in predicting house prices. This study will be of tremendous benefit, especially to housing developers and researchers, to ascertain the most significant attributes to determine house prices and to acknowledge the best machine learning model to be used to conduct a study in this field.

House Price Prediction (HPI) is commonly used to estimate the changes in housing price. As, house prices is dependent on many factors, like location, size, population, etc., it requires other information apart from HPI to predict individual house prices.

TABLE OF CONTENTS

1. INTRODUCTION

Development of civilization is the foundation of the increase in demand for houses day by day. Accurate prediction of house prices has been always a fascination for buyers, sellers, and bankers also. Many researchers have already worked to unravel the mysteries of the prediction of house prices. Many theories have been given birth as a consequence of the research work contributed by various researchers all over the world. Some of these theories believe that the geographical location and culture of a particular area determine how the home prices will increase or decrease whereas other schools of thought emphasize the socio-economic conditions that largely play behind these house price rises.

We all know that a house price is a number from some defined assortment, so obviously prediction of prices of houses is a regression task. To forecast house prices one person usually tries to locate similar properties in his or her neighborhood and based on collected data that person will try to predict the house price.

All these indicate that house price prediction is an emerging research area of regression that requires the knowledge of machine learning. This has motivated me to work in this domain.

Real estate appraisal is an integral part of the property buying process. Traditionally, the appraisal is performed by professional appraisers specially trained for real estate valuation. For the buyers of real estate properties, an automated price estimation system can be useful to estimate the prices of properties currently on the market. Such a system can be particularly helpful for novice buyers who are buying a property for the first time, with little to no experience.

The problem falls under the category of Supervised Learning algorithms. The dataset we'll be using, comprises of some input features and one target feature. The input features include features that may or may not impact the price.

2. RELATED WORK

[1]The paper studies the SVM algorithm in machine learning for house price prediction. It takes data from the user and process it and classify it using pre-available data and uses various classification algorithm and classifies data and predict the accurate price of the property.

[2]This paper proposed the FSTM to explore the spatiotemporal characteristics of the residential house price in Shunde as an example of the middle-small cities in China.

[3]The author constructs a fundamental algorithm based on the multiple linear regression method to predict housing prices and combines it with the Spearman correlation coefficient to determine the influential factors affecting housing prices.

[4]The study shows a comparison between the regression algorithms and artificial neural network when predicting house prices in Ames, Iowa, United States and Malmö, Sweden.

[5]This study is an exploratory attempt to use three machine learning algorithms in estimating housing prices, and then compare their results.

[6]This article concentrates on the comparison between different machine learning algorithms (Multiple Linear Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, Ada Boosting Regression, gradient boosting) about House price prediction Analysis.

[7]The study proposed a house prices prediction algorithm in Ames, lowa by deliberating on data processing, feature engineering and combination forecasting.

[8]This paper seeks useful models for house price prediction. It also provides insights into the Melbourne Housing Market. The evaluation phase indicates that the combination of Step-wise

and SVM model is a competitive approach.

[9]This paper proposes a hybrid Lasso and Gradient boosting regression model to predict individual house price. The proposed approach has recently been deployed as the key kernel for Kaggle Challenge “House Prices: Advanced Regression Techniques”.

[10]In this paper, SVM, LSSVM, and PLS algorithms are used in the field of construction to predict the housing value.

3. PROBLEM STATEMENT

We are given dataset of house prices with some features like no. of rooms, location, area, etc. Our task is to create a model which will predict the price for any new house by looking at its features.

The No Free Lunch Theorem state that algorithms perform differently when they are used under the same circumstances [2]. This study aims to analyze the accuracy of predicting house prices when using linear regression, Ridge, Lasso, Decision tree, Random forest, Elastic net, Bayesian Ridge. Thus, the purpose of this study is to deepen the knowledge in regression methods in machine learning. In addition, the given datasets should be processed to enhance performance, which is accomplished by identifying the necessary features by applying one of the selection methods 2 to eliminate the unwanted variables since each house has its unique features that help to estimate its price. These features may or may not be shared with all houses, which means they do not have the same influence on the house pricing resulting in inaccurate output.

4. OBJECTIVE

Following are the objectives of the project:-

Predict the price of any new house based on its feature (location, area, no. of rooms), which will include following sub objectives:

- Finding the type of model to build
- Selecting a performance measure
- Checking assumptions

5. DESIGN

5.1 Methodology

We will use a test driven approach to build a model using Python and then we will use trained model to predict house sale prices and extend it further.

Steps:

- Import dataset and necessary libraries
- Data preprocessing
- EDA
- Data cleaning: It involves dealing with missing values and outliers, dimensionality reduction
  - Build model: using linear regression, Ridge, Lasso, Decision tree, Random forest , Elastic net, Bayesian Ridge
  - Compare accuracy: as Linear Regression and Bayesian regression is giving more accurate results, so use that for testing
  - Test the model for certain parameters

5.2 Algorithm

1. **Linear Regression**

Simple linear regression uses a traditional slope-intercept form, where a and b are the coefficients that we try to “learn” and produce the most accurate predictions. X represents our input data and Y is our prediction.

Y=bX+a
Y’= A+B*X

Where X: Predictor (present in data)
B: coefficient (estimated by regression)
A: intercept (Estimated by regression)
Y’: Predicted value (calculate from A, B and X.

Multivariable Regression

A more complex, multi-variable linear equation might look like this, where w represents the coefficients or weights, our model will try to learn.

Y(x1, x2, x3)=w1x1+w2x2+w3x3+w0

The variables x_1, x_2, x_3 represent the attributes or distinct pieces of information, we have about each observation.

Loss function

Given our Simple Linear Regression equation:

Y=BX+a
We can use the following cost function to find the coefficients/parameters for our model:
Mean Squared Error (MSE) Cost Function
The MSE is defined as:

$$MSE = J(W) = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h_w(x^{(i)}))^2$$

where

$$h_w(x) = g(w^T x)$$

The MSE measures how much the average model predictions vary from the correct values. The number is higher when the model is performing "bad" on our training data.
The first derivative of MSE is given by:

$$MSE' = J'(W) = \frac{2}{m}\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})$$

One Half Mean Squared Error (OHMSE)
We will apply a small modification to the MSE — multiply by 1/2 so when we take the derivative, the 2s cancel out:

$$OHMSE = J(W) = \frac{1}{2m}\sum_{i=1}^{m}(y^{(i)} - h_w(x^{(i)}))^2$$

The first derivative of OHMSE is given by:

$$OHMSE' = J'(W) = \frac{1}{m}\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})$$

2. **Ridge regression**

- Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.
- Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as L2 regularization**.**
- In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called Ridge Regression penalty**.** We can calculate it by multiplying with the lambda to the squared weight of each individual feature.
- The equation for the cost function in ridge regression will be:
- $$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}\beta_j^2$$
- In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the amplitudes of the coefficients that decreases the complexity of the model.
- As we can see from the above equation, if the values of λ tend to zero, the equation becomes the cost function of the linear regression model. Hence, for the minimum value of λ, the model will resemble the linear regression model.

- A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.
- It helps to solve the problems if we have more parameters than samples.

3. Lasso Regression

- Lasso regression is another regularization technique to reduce the complexity of the model. It stands for Least Absolute and Selection Operator.
- It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.
- Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
- It is also called as L1 regularization. The equation for the cost function of Lasso regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}|\beta_j|$$

- Some of the features in this technique are completely neglected for model evaluation.
- Hence, the Lasso regression can help us to reduce the overfitting in the model as well as the feature selection.

4. Decision Tree

The decision tree is an important algorithm for predictive modelling and can be used to visually and explicitly represent decisions. It is a graphical representation that makes use of branching methodology to exemplify all possible outcomes based on certain conditions. In decision tree internal node represents a test on the attribute, branch depicts the outcome and leaf represents decision made after computing attribute. Decision Tree helps in making decisions under a particular circumstance and improves communication. It helps data scientist to capture the idea that how different decisions can lead to different operational nature of the situation. It helps to take an optimal decision. The algorithm is well suited for problems where instances are represented by attribute value and when training data contains error. It is also applicable to the situation when the target function has discrete output value.

This model called Decision tree and we will make prediction based on this model. It divides houses into only two categories. The predicted price for any house under consideration is the historical average price of houses in the same category. We use data to decide how to break the houses into two groups, and then again to determine the predicted price in each group. This step of capturing patterns from data is called fitting or training the model. The data used to fit the model is called the training data. There are too many variables to wrap your head and you can not distinguish which is the most relevant one. So at first, we start by picking a few varibles using our intuition. Next, we will choose prediction target and relevant features. Choose Prediction Target: We'll use the dot notation to select the column we want to predict, which is called the prediction target. By convention, the prediction target is called y. Choose ―Features‖: The columns that are inputted into our model (and later used to make predictions) are called ―features.‖ In our case, those would be the columns used to determine the home price. Sometimes, you will use all columns except the target as features. Other times you'll be better off with fewer features. For now, we'll build a model with only a few features.

5. Random forest

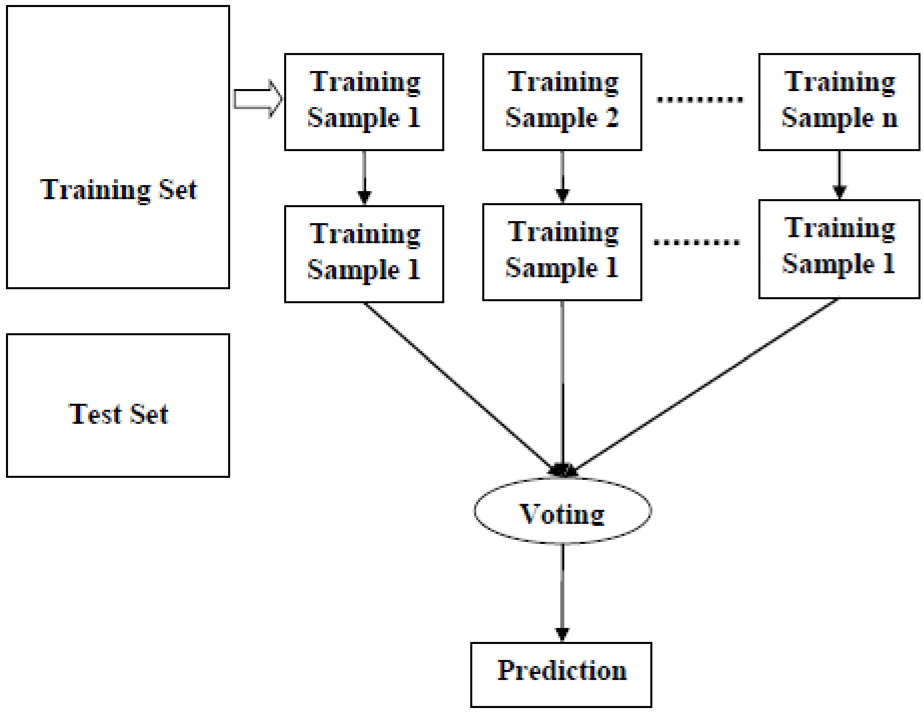
Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps −

- **Step 1** − First, start with the selection of random samples from a given dataset.
- **Step 2** − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- **Step 3** − In this step, voting will be performed for every predicted result.
- **Step 4** − At last, select the most voted prediction result as the final prediction result.

- The following diagram will illustrate its working −



-

6. Elastic net

Elastic-Net Regression is a modification of Linear Regression which shares the same hypothetical function for prediction. The cost function of Linear Regression is represented by *J*.

$$\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}-h\left(x^{(i)}\right)\right)^2$$

Here, m is the total number of training examples in the dataset.

$h(x^{(i)})$ represents the hypothetical function for prediction.
$y^{(i)}$ represents the value of target variable for i^{th} training example.
Linear Regression suffers from overfitting and can't deal with collinear data. When there are many features in the dataset and even some of them are not relevant for the predictive model. This makes the model more complex with a too inaccurate prediction on the test set (or overfitting). Such a model with high variance does not generalize on the new data. So, to deal with these issues, we include both L-2 and L-1 norm regularization to get the benefits of both Ridge and Lasso at the same time. The resultant model has better predictive power than Lasso. It performs feature selection and also makes the hypothesis simpler. The modified cost function for Elastic-Net Regression is given below :

$$\frac{1}{m}\left[\sum_{l=1}^{m}\left(y^{(i)}-h\left(x^{(i)}\right)\right)^2+\lambda_1\sum_{j=1}^{n}w_j+\lambda_2\sum_{j=1}^{n}w_j^2\right]$$

Here, $w_{(j)}$ represents the weight for j^{th} feature.
n is the number of features in the dataset.
lambda1 is the regularization strength for L-1 norm.
lambda2 is the regularization strength for L-2 norm.

7. **Bayesian Ridge**

   Bayesian regression allows a natural mechanism to survive insufficient data or poorly distributed data by formulating linear regression using probability distributors rather than point estimates. The output or response 'y' is assumed to drawn from a probability distribution rather than estimated as a single value.

   Mathematically, to obtain a fully probabilistic model the response y is assumed to be Gaussian distributed around XwXwXas follows
   p(y|X,w,α)=N(y|Xw,α)

   One of the most useful type of Bayesian regression is Bayesian Ridge regression which estimates a probabilistic model of the regression problem. Here the prior for the coefficient w is given by spherical Gaussian as follows –

   p(w|λ)=N(w|0,λ−1Ip)

8. **Outliers**

Outliers are noisy data that they do have abnormal behaviour comparing with the rest of the data in the same dataset. Outliers can influence the prediction model and performance due to its oddity. There are three types of outliers, which are point, contextual, and collective outliers [29]. Point outlier is an individual data instance that can be considered as odd with respect to 15 the rest of the data. The contextual outlier is an instance of data that can be regarded as odd in a specific context but not otherwise. An example of contextual is the longitude of a location. A collective outlier is a collection of related data instances that can be considered as abnormal with respect to the entire dataset. In supervised, the detection of outliers can be accomplished visually, where a predictive model is built for normal against outliers' classes. Dean De has investigated the public dataset and he suggests to remove certain outliers from the public data when he said "I would recommend removing any houses with more than 4000 square feet from the data set" [30]. Another example of detecting outliers is by using Isolation forest, which has two stages, training, and testing. The training is to create the isolation trees and then to record the anomaly score of each entry in the testing stage. This method has shown a promising result, according to [31].
As a data scientist when you have a conversation with your business manager (who has expertise in real estate), he will tell you that normally square ft per bedroom is 300 (i.e. 2 bhk apartment is minimum 600 sqft. If you have for example 400 sqft apartment with 2 bhk than that seems

suspicious and can be removed as an outlier. We will remove such outliers by keeping our minimum thresold per bhk to be 300 sqft.

Outlier Removal Using Standard Deviation and Mean

Another way we can remove outliers is by calculating upper boundary and lower boundary by taking 3 standard deviation from the mean of the values (assuming the data is Normally/Gaussian distributed).

Use One Hot Encoding For Location

One-hot encoding is a technique that is used to convert categorical features to a suitable format to be used as an input in Machine Learning algorithms [34]. It transforms a single variable with n observations and d distinct values to d binary variables, where each observation indicating the presence as 1 or absence as 0 [35]. In one-hot encoding, the categories are represented as independent concepts.

6. SWOT Analysis

6.1 Strength

1. Simple model
2. Computationally efficient
3. Interpretability of the output

6.2 Weakness

1. Overly-simplistic
2. Independence of variables
3. Inability to determine feature importance

6.3 Opportunities

1. Data security improvement
2. More accessible
3.

6.4 Threats

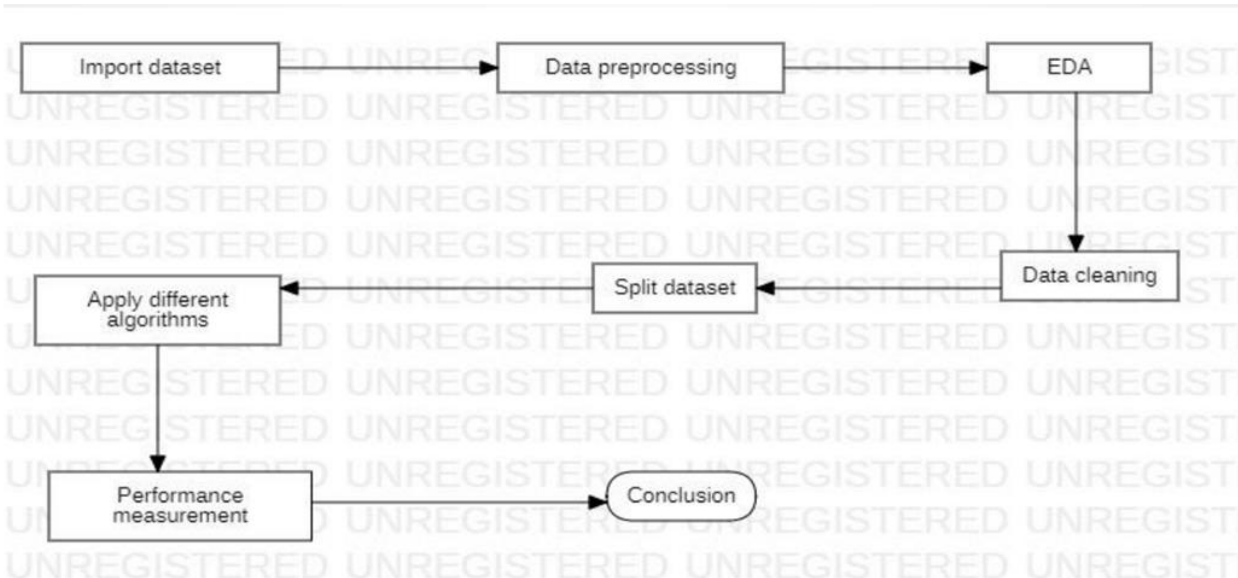
1. If data cleaning is not done efficiently, then the outcome would not be as desired.

7. Applications

1. Government Sectors (Statistics)
2. Corporate Sectors
3. Real estate
4. Investors

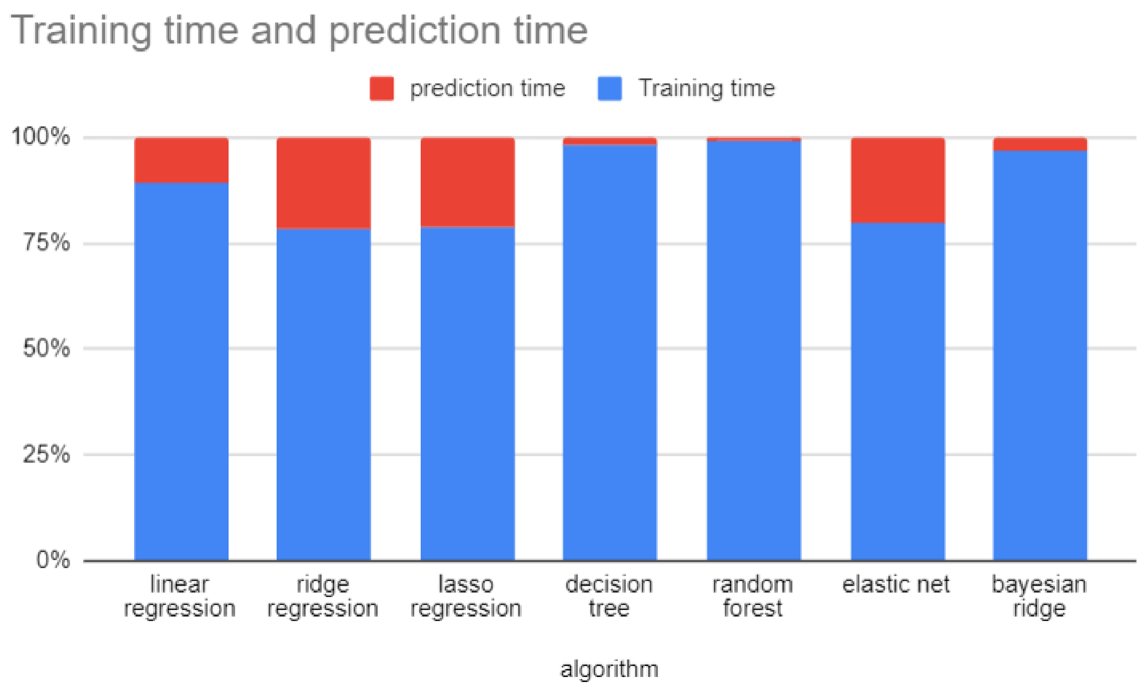
8. Implementation

8.1 Flow Chart

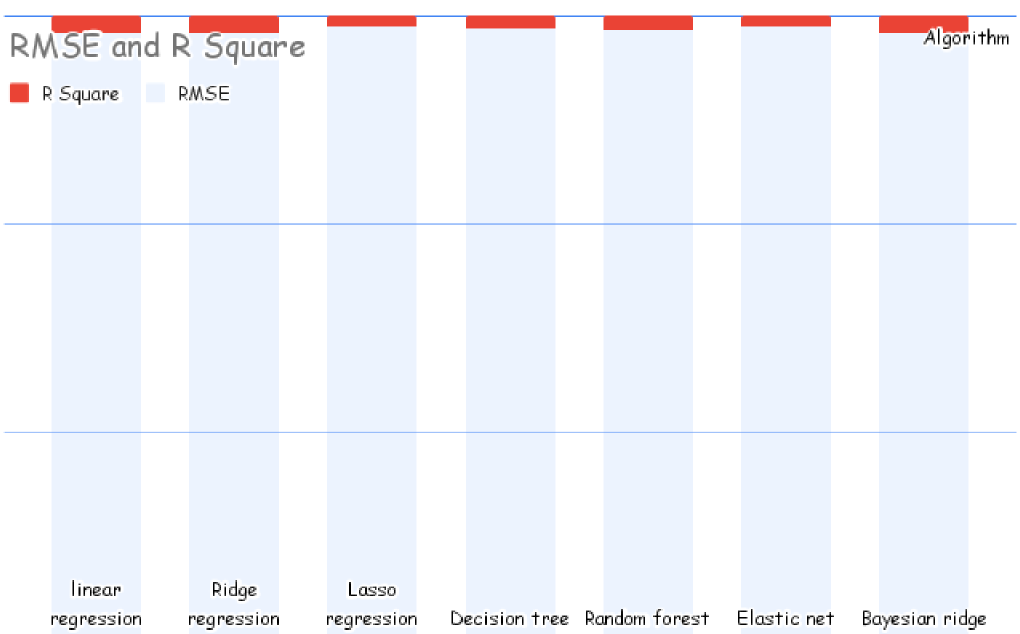


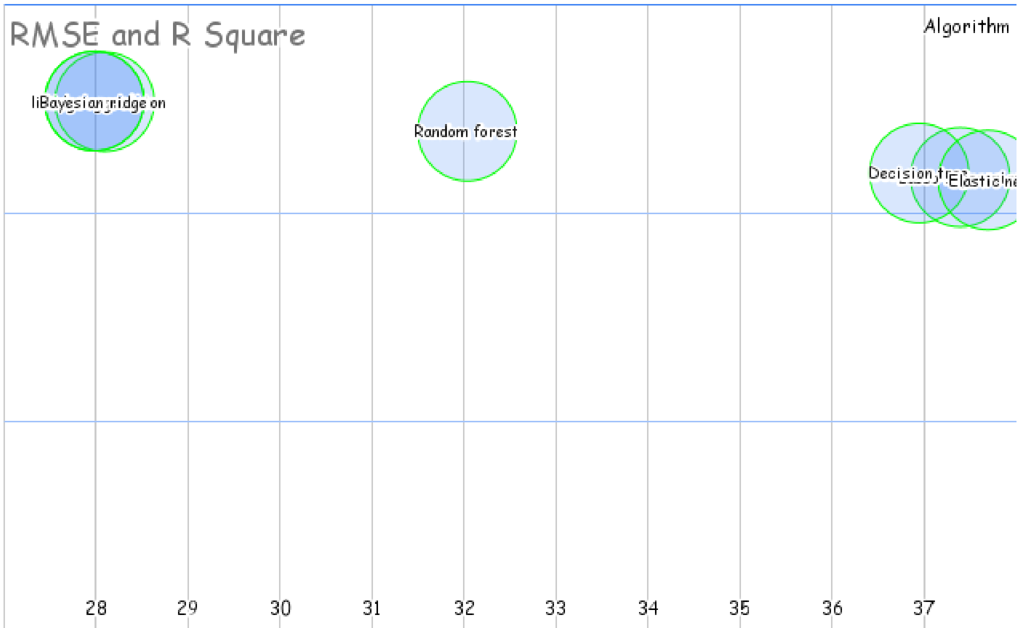
9. Result and Analysis

Algorithm	Training time(s)	Prediction time(s)	Explained variance	Mean absolute error	Mean squared error	RMSE	R2 square
Linear Regression	0.066	0.008	0.845498	16.59489	783.26599	27.98688	0.84522
Ridge Regression	0.029	0.008	0.844342	16.3832	789.5313	28.0986	0.84398
Lasso Regression	0.030	0.008	0.7248814	22.1792	1397.8972	37.3884	0.72377
Decision tree	0.248	0.005	0.736080	19.70818	1364.7815	36.94295	0.73032
Random forest	6.288	0.046	0.80386	17.6211	1026.4229	32.03783	0.79718
Elastic net	0.016	0.004	0.72046	22.2640	1420.4937	37.6894	0.71931
Bayesian Ridge	0.186	0.006	0.84553	16.4414	783.3193	27.9878	0.84521



Training time for Random forest is maximum as compare to other algorithm and the prediction time for Elastic Net is minimum.





The best algorithm among the above algorithm is that whose R^2 value is maximum and the RMSE value is minimum, it lied upper left corner of the graph. And it is Bayesian ridge and linear regression.

10. CONCLUSION AND FUTURE SCOPE

10.1 Conclusion

Taking the sample dataset for houses, and considering its various attributes, the prices for houses have been predicted by employing machine learning methods of regression-for predicting the price of estate using prior data, and clustering- for inspecting the quality of the solution or output. Also, multi collinearity removal using VIF has been applied for removing the errors and redundancy from the dataset using Python.

Linear Regression and Bayesian ridge algorithm gives the best result.

Bayesian ridge is:

- Very effective when the size of the dataset is small.
- Particularly well-suited for on-line based learning (data is received in real-time), as compared to batch based learning, where we have the entire dataset on our hands before we start training the model. This is because Bayesian Regression doesn't need to store data.
- The Bayesian approach is a tried and tested approach and is very robust, mathematically. So, one can use this without having any extra prior knowledge about the dataset.

Linear Regression is:

- Linear Regression is simple to implement and easier to interpret the output coefficients.
- When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of it's less complexity to compared to other algorithms.
- Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

10.2 Future Scope

The data modelling and analysis in this work has scope for future application in lodging value-prediction systems. Based on the results, it can be concluded that such ML-driven predictions are easily comprehendible and significant from a data-analytics point of view. When correctly implemented, a high rate of accuracy can be achieved, and thus ML techniques find applications across a wide range of fields

11. References

1. Mr. Piyush Chordia, Mr. Pratik Konde, Ms. Supriya Jadhav, Hrutik Pandhare, Prof. Shikha Pachouly. Prediction of House Price Using Machine Learning 2022. https://doi.org/10.22214/ijraset.2022.40466
2. Lu Wang,Guangxing Wang,Huan YuORCID Icon &Fei Wang. Prediction and analysis of residential house price using a flexible spatiotemporal model 2022 https://www.tandfonline.com/doi/full/10.1080/15140326.2022.2045466
3. Qingqi Zhang. Housing Price Prediction Based on Multiple Linear Regression 2021 https://doi.org/10.1155/2021/7678931
4. Ahmad Abdulal Nawar Aghi. House Price Prediction 2020 https://www.divaportal.org/smash/get/diva2:1456610/FULLTEXT01.pdf
5. Winky K.O. Ho,Bo-Sin Tang &Siu Wai Wong. Predicting property prices with machine learning algorithms 2020 https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558
6. CH. Raga Madhuri; G. Anuradha; M. Vani Pujitha. House Price Prediction using Regression Techniques: A Comparative Study 2019 DOI: 10.1109/ICSSS.2019.8882834
7. Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing ICMLC 2018. doi:10.1145/3195106.3195133.
8. Phan TD. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (ICMLDE) 2018. doi:10.1109/icmlde.2018.00017.
9. Lu S, Li Z, Qin Z, Yang X, Goh RSM. A hybrid regression technique for house prices prediction. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) 2017. doi:10.1109/ieem.2017.8289904.
10. Mu J, Wu F, Zhang AHousing Value Forecasting Based on Machine Learning Method**s**Abstract and Applied Analysis, 2014 (2014), pp. 1-7doi:10.1155/2014/648047.