House Price Prediction Using Regression Techniques

Frency Chauhan, Nirva Sangani, Yagnik Hingrajiya, Vijay Rana
Ahmedabad University, frency.c, nirva.s, yagnik.h, vijay.r@ahduni.edu.in

***Abstract* - The decision making for investment in real estate can be a task with lots of thinking. With the frequent fluctuating prices and fulfilling the expectation of the investor, there arises the need for a mathematical model that helps the user with their decision making process. In this work, the house price prediction is done by using different machine learning techniques like linear regression, gradient boosting regression, polynomial regression and random forest regression algorithm and comparison of mean square error of linear regression, ridge regression and lasso regression.**

Key Words - House Price, Prediction, Linear Regression, Maximum Likelihood Estimation, Gradient Boosting, Polynomial Regression, Random Forest Regression, Mean Square Error, Chi Square Test, Feature Selection, Ridge Regression, Lasso Regression

I. INTRODUCTION

With the increase in population the demand for buying houses is increasing as well. The price prediction for a house depends on various parameters like the location of the house, area, number of bedrooms etc. As price varies with these parameters, a challenge to make a model that accurately predicts the price emerges. This paper considers the problem of predicting house price for the houses of Mumbai city as a regression problem and explores various regression techniques like linear regression with maximum likelihood estimation and gradient boosting, polynomial regression, ridge regression, lasso regression and random forest regression algorithm to predict the price of house in Mumbai city.

II. LITERATURE SURVEY

Predicting the price of the house is the problem that has been tackled both as a regression problem and a classification problem by different researchers.

Manasa, Gupta and Narahari [2] have used house price data of Bengaluru city and predicted price by using various regression techniques such as Lasso regression, Ridge regression, Support vector regression (SVR) and Extreme gradient boosting (XGBoost) regression and have used the evaluation metrics as root mean square error (RMSE), R-square and adjusted R-square and root mean squared logarithmic error (RMSLE).

Sawant, Jangid, Tiwari, Jain and Gupta [3] have used house price data of Pune city and predicted price by using decision tree and random forest algorithms.

Durganjali and Pujitha [4] have used various classification algorithms like Logistic regression, Decision tree, Naive Bayes and Random forest.

Varma, Sarma, Doshi and Nair [5] have used Linear regression, random forest regression, boosting regression algorithm and neural networks approach to predict the house price.

III. IMPLEMENTATION

1. DATA SET: We have acquired house price data of Mumbai city from Kaggle website [1]. The data set consists of different parameters like price of the house, area in square feet, locality, number of bedrooms and bathrooms etc.

For all the models used except ridge and lasso regression, the evaluation metric for train and test data is accuracy in percent. For ridge and lasso regression, the evaluation metric is mean square error.

TABLE 1: Specifications of data set

Number of columns	19
Number of rows	6348
Type of problem	Regression
Target variable	Price
Missing of value	NIL
Choice of evaluation metrics	Accuracy in percent, Mean Square Error

2. FLOW OF IMPLEMENTATION OF EVERY MODEL:

I. Data acquisition
II. Data pre-processing
III. Model training
IV. Model evaluation
V. Hyperparameter tuning
VI. Model Deployment

3. FEATURE SELECTION: As the dataset contains categorical data, the feature selection has been done using Chi Square Test and top five features are selected.

The Chi Square score is calculated as:

X^2 = *(Observed Frequency - Expected Frequency)^2 / Expected Frequency*

The five features from the dataset are obtained using Chi square test and the regression models are implemented on these selected features.

Selected five features are:

I. Area
II. Number of Bedrooms
III. Landscaped gardens
IV. Indoor games
V. Gas connection

4. MODELING: The different machine learning based models used to predict the price of the house are:

- Linear regression with gradient boosting
- Linear regression with maximum likelihood estimation
- Polynomial regression (degree = 2)
- Random forest regression
- Linear regression model from scratch
- Ridge regression
- Lasso regression

I. LINEAR REGRESSION:

$$Y = \Theta_0 + \Theta_1 X_1 + +\Theta_2 X_2 + \cdots$$

$\theta_i = parameter$

$x_i = feature$

II. RIDGE REGRESSION:

$$J(\theta) = \min_{\theta} \frac{1}{2m}[\sum_{i=1}^{m} h_\theta(x^{(i)} - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2]$$

$$\Theta_j := \Theta_j(1 - \propto \frac{\lambda}{m}) - \frac{\propto}{m}\sum_{i=1}^{m}(\boldsymbol{x^{(i)}} - \boldsymbol{y^{(i)}})\,\boldsymbol{x_j^{(i)}}$$

III. LASSO REGRESSION:

$$J(\theta) = \min_{\theta} \frac{1}{2m}[\sum_{i=1}^{m} h_\theta(x^{(i)} - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|]$$

$$\Theta_j := \Theta_j(1 - \propto \frac{\lambda}{m}) - \frac{\propto}{m}\sum_{i=1}^{m}(\boldsymbol{x^{(i)}} - \boldsymbol{y^{(i)}})\,\boldsymbol{x_j^{(i)}}$$

IV. MEAN SQUARE ERROR:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

N= number of examples

IV. RESULTS

TABLE 2: Accuracy Results of train and test data set for different techniques used

Model	**Train Data Accuracy**	**Test Data Accuracy**
Linear Regression(Library)	54%	47%
Linear Regression with Maximum Likelihood Estimation(Scratch)	54%	47%
Gradient Boosting Regression	93%	49%
Polynomial Regression	64%	63%

Random Forest Regression	92%	54%

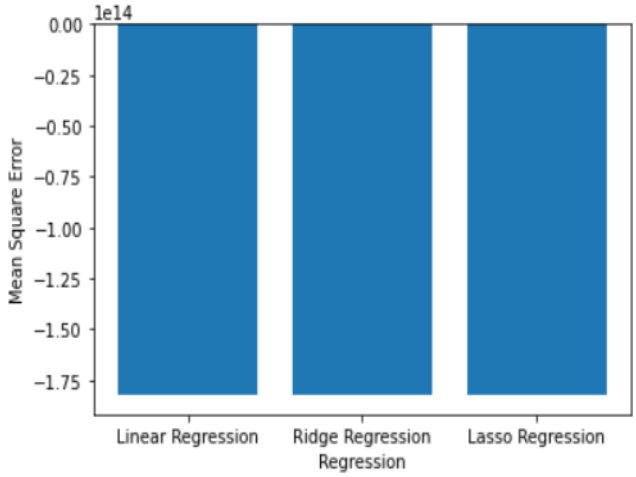

Figure 1: Mean Square Error comparison between linear regression, ridge regression and lasso regression models

V. Conclusion

Observing table 2, it is evident that gradient boosting technique for linear regression gives higher accuracy than simple linear regression and maximum likelihood estimation for linear regression. The random forest regression gives almost the same accuracy as gradient boosting for train data but both the models give comparatively lower accuracy for test data. This is happening due to the overfitting problem. Polynomial regression gives higher accuracy than simple linear regression and linear regression with maximum likelihood estimation. Accuracy for linear regression implementation using built-in libraries and linear regression implementation from scratch - both gives exact same accuracies which shows the correctness of implemented model from scratch. From the bar graph, it is evident that mean square errors for linear regression, ridge regression and lasso regression do not show significant variation from each other, hence for this dataset used, all three regression models fit almost identically.

VI. References

[1] https://www.kaggle.com/sameep98/housing-prices-in-mumbai

[2] J. Manasa, R. Gupta and N. S. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 624-630, doi: 10.1109/ICIMIA48430.2020.9074952.

[3] R. Sawant, Y. Jangid, T. Tiwari, S. Jain and A. Gupta, "Comprehensive Analysis of Housing Price Prediction in Pune Using Multi-Featured Random Forest Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697402.

[4] P. Durganjali and M. V. Pujitha, "House Resale Price Prediction Using Classification Algorithms," 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2019, pp. 1-4, doi: 10.1109/ICSSS.2019.8882842.

[5] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.