A Futuristic Linear Regression Model for Housing

K.S. Seetha Raman[#1], B.B.A. Amrutha[#2], N. Naveena[#3], P. Nivetha[#4], S. Malathi[#5]

[1]*Assistant professor,Department of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India*

[2]*Final year student, Department of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India*

[3]*Final year student, Department of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India*

[4]*Final year student, Department of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India*

[5]*Final year student, Department of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India*

Abstract

In this paper, we are predicting the price of houses based on the several factors or features found in the data set such as crime rate, population, age, tax rate, pupil-teacher ratio, distance from employment centers, average number of rooms, accessibility to highways and so on. Generally, house prices are determined by factors like geographical location, size of the plot, population, cost of living in that country, poverty, people's income, number of bedrooms in the house etc. It helps buyers as well as the sellers to make informed decisions. Hence it is good to have a system that defines the relationship between real estate or house prices and the various influencing factors. This paper uses Linear Regression technique for predicting the price. Linear regression is a supervised machine learning algorithm that designs a model that best suits the given dataset to predict a dependent variable from a set of explanatory variables. We use feature selection which gives a score on how much each feature affects the dependent variable. We will be considering a data set called Boston Housing Prices Dataset. It has 506 instances and 13 attributes along with one target variable. Initially, in the preprocessing stage, data will be made fit for regression. Then the linear regression model is applied. As a result, we obtain the regression parameters. Using these parameters, we write a linear equation. This is called the Linear Regression Model. Prediction error is calculated. The objective of the project is to minimize this error as much as possible and deliver an end product that can be used by the common public to help predict the price of their property. Record theerror or accuracy obtained. Upon training the model with data from previous years, it will be fit to predict the futuristic prices or for the features fed by the users, who may be both buyers and sellers.

Key words - *Linear Regression, Multiple Linear Regression, Real Estate, House price prediction, Feature selection, Supervised Machine Learning*

I.INTRODUCTION

House prices have always attracted the attention of real estate developers, banks, policy makers or, in short, the general public as well as to actual and potential home owners. Valuations of housing are necessary in order to assess the benefit and liabilities in housing section. These valuations are performed by the different players in the marketplace such as real estate agents, appraisers, assessors, mortgage lenders, brokers, property developers, investors and fund managers, lenders, market researchers and analysts and other specialists and consultants.

A diversity of methods that are utilized in estimating the value in literature is at hand. Thus, predicting house price, which is a continuous valued output variable uses regression technique. If we predict more than one output variable, it is called Multi target regression. Since we predict only one output variable namely the price of the house, it is called Linear regression. More specifically, it is Multiple Linear Regression. Multiple regression model is one that predicts a dependent variable which is the output variable, based on the value of two or more independent variables. This project uses Linear Regression model for the prediction.

In statistics, linear regression is a techniquefor establishing a relationship betweendependentvariable

and one or moreexplanatoryvariablesalso called as the independentvariables. The dependent variable takes scalar values. A simplelinearregression is one where there is exactly one dependent variable . If there are more than one input features, it is termed as multiple linear regression. This term is different from multivariatelinearregression, where multiple scaloutput variables are predicted.In linear regression, linear predictor functions are used to model the relationships.

The main objective of this project is to deliver an end product that receives input from user, use them to calculate the price of the house and give them the price predicted by the linear regression model. We create a web static web page that has text areas to receive input, a button to invoke a function calculating the price and display the calculated result. The data set should be trained and as a result of applying and fitting the regression model, we get the regression parameters also called as the regression coefficients.

I.LITERATURE SURVEY

Ceyhun Ozgur [1] has implemented multiple linear regression on a housing data set. Scatter plot and box plot have been used to visualize the graph. For scalar values, we use scatter plot. Initially, mean, median and standard deviation for the data are calculated. Correlation between variables are plotted. Pearson's correlation coefficients are tabulated. Correlation refers to the statistical measure that indicates the extent to which two or more variablesdepend on or affect each other. A good correlation analysis generally results to a better understanding of our dataset. Then, scatter plots between the target variable, price vs hoa, price vs size, price vs yards, price vs age, price vs floors, price vs bedrooms and price vs bathrooms. Boxplot with the same features as x and y axes are plotted. Parameter estimates are done and standard errors of the estimates, including t-tests and p-values are obtained. Based on their results, one feature 'hoa' is picked to be significant in relation to the output. This paper analyzes what factors would have influenced the price of the houses. Then, we chose seven variables namely HOA, size, age, yards, floors, bedrooms and bathrooms to include in our model. After further analysis we finally include HOA in our model. The easiest model is considered the best and so we chose to use one variable to for prediction. Finally, the equation: **Price=312638+17.854Hoa**is taken as the regression model.

In [2], we use various machine learning algorithms to predict the house prices. Numerous factors such as area of the property, location of the house, material used for construction, age of the property, number of bedrooms and garages and so on affect the house price. Here, machine learning algorithms such as Logistic regression, Support Vector Regression, Lasso Regression and Decision Tree are employed to build predictive models.Logistic Regression, SVM, Lasso Regression and Decision Tree result in the R-squared value of 0.98, 0.96,0.81 and 0.99 respectively. Other error calculation methodologies like MAE, MSE and RMSE have been calculated and the algorithms are compared based on their values.

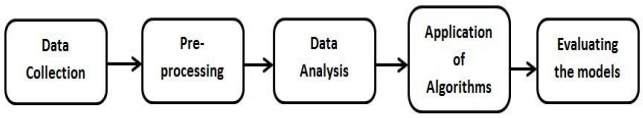

Fig 1: System Architecture

This paper also proposes a neat system architecture for the implementation of the project. The steps involved are Data Collection, Pre-processing, Data Analysis, Application of algorithms and the final stage being evaluating the models.

In [3], data source on average price and influencing factors of China's housing sales from 2000 to 2015 from China statistical yearbook is taken. 16 years of data were analyzed in time series and multiple regression models were set up. Then we check the stability of the data by performing ADF (Augmented Dickey-Fuller) test. Correlation coefficient matrix is generated. Variation Inflation Factor is calculated for each explanatory variable. It is concluded by stating the changes in percentage of the output variable, the price of the house with respect to changes in each explanatory variable.

The paper [4] initially classifies the attributes into three groups namely physical condition, concept and location. Hedonic pricing is used to calculate the price of the house. It refers to the prediction model based on the hedonic pricing theory. The theory states that the value of the property is the sum of all its attribute values. Regression analysis and particle swarm optimization are the two major methodologies in this paper. The dataset is taken and after preprocessing it is ready to perform regression analysis. We obtain the regression equation.As a result of applying these algorithms, error values based on MAPE, RMSE and MAE are calculated.

Hedonic theory states that the price of a commodity is the function of all of its attributes. Hence the price of house can be derived from its explanatory variables available in the dataset. An introduction about linear regression and regression tree are given. It uses Mutual

Information matrix to measure the statistical relationship between any two variables. VIF (Variance Inflation Factor) states the absence of collinearity. Both linear regression and regression tree model are built and the outcomes are tabulated in [5].

In [6], Feature selection is done initially. A partial category hierarchy of house features are shown. Topic modelling and deep learning are used for feature selection. Event and time, Right censoring and Proportional hazard model are the primary components of the Survival Regression model. Data description, data clean up, data statistics are considered in data section. Histograms are used for visualizing. The experimental setup involves the implementation of the algorithms. Concordance index is used as the evaluation metric. After performance analysis, heat maps are generated.

Fuzzy linear regression and artificial neural network are used here [7]. In multilayer perceptron, the data flows forward towards the last output layer without feedback. Error back propagation algorithm adjusts the weight in the artificial neural network in order to reduce the error. The results indicate that the estimation error the ANN-Back propagation technique is less than that in Fuzzy regression.

II.EXISTING SYSTEM

The existing systems predict the futuristic prices based on only one best feature taken into considerations. This, however being a simple model, does not seem to be the best or accurate model for predicting house prices. Various data visualization techniques are used to graphically visualize the dataset. Various error calculation methodologies are used to calculate the rate of error. In machine learning algorithms, error is taken as the difference between the recorded or observed values in the dataset and the values predicted by our prediction algorithm.

III.PROPOSED SYSTEM AND IMPLEMENTATION

The workflow diagram is shown below.

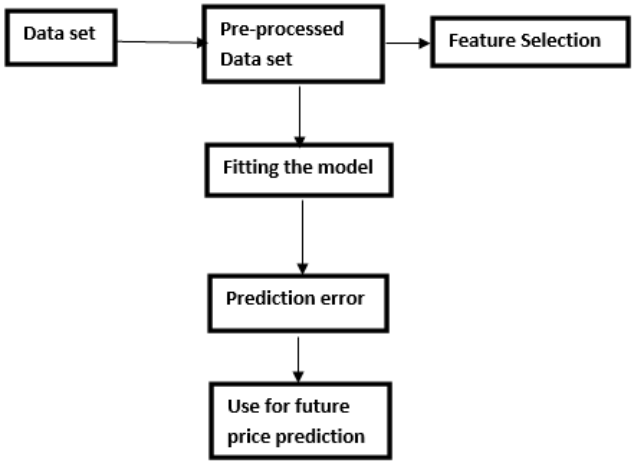

Fig 2: Work flow

The modules of the linear regression model and their description are as follows.

1) ***Dataset Collection:***

Searching for 'the perfect dataset' is an important part. After going through the features, the Boston housing prices dataset [9] is selected. It has 506 instances, 13 attributes and one target variable namely the observed price of the house. There are no missing values or null values in the dataset. The following are the input features and their explanations present in the dataset.

TABLE II

ATTRIBUTES, TARGET VARIBLE AND THEIR EXPLANATION

FEATURE	EXPLANATION
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq. ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distance to five Boston employment centers
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63) ^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

*2)**Pre-processing:***

Importing all the packages necessary to fit and train the linear regression model. Load the comma separated values into Pandas DataFrame. Add the target variable to the DataFrame.

*3)**Fitting the model – Regression coefficients***:

Find the regression coefficients also called as the regression parameters. Fit the model using LinearRegression method from sklearn. The method 'fit' takes two parameters, x being the input variable and y, the output variable. Obtain the 13 regression coefficients and one intercept values.

Intercept value = 36.459488385089855

TABLE II

REGRESSION PARAMETERS FOR EACH ATTRIBUTES

FEATURE	REGRESSION COEFFICIENT
CRIM	-0.108011
ZN	0.046420
INDUS	0.020559
CHAS	2.686734
NOX	-17.766611
RM	3.809865
AGE	0.000692
DIS	-1.475567
RAD	0.306049
TAX	-0.012335
PTRATIO	-0.952747
B	0.009312
LSTAT	-0.524758

*4) **Prediction error calculation***:

RMSE is used to predict the error. RMSE is Root Mean Squared Error. It is the measure of differences between the values predicted by a model or estimator and the values observed. The formula for RMSE is given as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

*5)**Feature Selection:***

Feature selection has other names such as variable selection, attribute selection and variable subset selection. It is the process of calculating a score of what measure each explanatory variable affects the output variable. The score is tabulated.

TABLE III

FEATURE SELECTION SCORES FOR EACH ATTRIBUTES

CRIM	0.112218
ZN	0.02277125
INDUS	0.05034649
CHAS	0.01497713
NOX	0.07242233
RM	0.13713406
AGE	0.12285762
DIS	0.10909201
RAD	0.03256386
TAX	0.04648173
PTRATIO	0.04317475
B	0.10218954
LSTAT	0.13377124

*6)**Predict prices:***

Use the prediction parameters to predict house prices for future data or data user enters. Let the data user enter be x1, x2, x3, x4, …., x13, then form a linear equation with the obtained parameters as follows.

Price predicted = 36.459488385089855-0.108011*x1+0.046420*x2+0.020559*x3+2.686734*x4-17.766611*x5+3.809865*x6+0.000692*x7-1.475567*x8+0.306049*x9-0.012335*x10-0.952747*x11+0.009312*x12-0.524758*x13

IV.CONCLUSION

Thus we can estimate the price of housing provided the given data. House prices keep changing every year. We can predict which factors strongly affect the price. It is useful to make informed decisions for buyers and sellers.

REFERENCES

[1] Ceyhun Ozgur, Ph.D., Zachariah Hughes, Grace Rogers, Sufia Parveen, "Multiple Linear Regression Applications in Real Estate Pricing", International Journal of Mathematics and Statistics Inventions(IJMSI), E-ISSN: 2321-4767 P-ISSN: 2321-4759 www.ijmsi.org Volume 4 Issue 8, October 2016

[2] Neelam Shinde, Kiran Gawande, "Valuation Of House Prices Using Predictive Techniques", International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-
5, Issue-6, Jun.-2018

[3] Wang Aiyin, Xu Yanmei, "Muliple Linear Regression Analysis of Real Estate Price", 2018
International Conference on Robots & Intelligent
System, 2018

[4] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, "Modeling house price prediction using regression analysis and particle swarm optimization Case study: Malang, East Java, Indonesia",IJACSA International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017.

[5] Timothy Oladunni, Sharad Sharma, "Spatial Dependency and Hedonic Housing Regression Model", 15th IEEE International Conference on Machine Learning and Applications, 2016.

[6] Mansurul Bhuiyan, Mohammad Al Hasan, "Waiting to be Sold: Prediction of TimeDependent House Selling Probability", 2016 IEEE International Conference on Data Science and Advanced Analytics, 2016.

[7] Reza Ghodsi, Abtin Boostani, Farshid Faghihi, "Estimation of Housing Prices by Fuzzy Regression and Artificial Neural Networks", 2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation, 2010.

[8] Boston house prices datsaset from https://scikitlearn.org/stable/modules/generated/sklearn.datasets.load_boston.html#sklearn.datasets.load_boston

[9] Towards Data Science article 'Create a model to predict house prices using Python' by Shreyas Raghavan https://towardsdatascience.com/create-amodel-to-predict-house-prices-using-pythond34fe8fad88f

[10] Data Science Plus article 'Linear Regression in Python' by Susan Li, Senior Data Scientist, Kognitiv Corp https://datascienceplus.com/linearregression-in-python-predict-the-bay-areas-homeprices/

[11] Medium article 'Predicting House Prices Using Linear Regression' by Gerald Muriuki https://medium.com/africa-creates/predictinghouse-prices-using-linear-regressionfe699d091e04

[12] Towards Data Science article 'Feature selection techniques in ML with Py' by Raheil Shaik https://towardsdatascience.com/feature-selectiontechniques-in-machine-learning-with-pythonf24e7da3f36e