Grenze International Journal of Engineering and Technology, Special Issue

Survey on House Price Prediction and Analysis

Nimish Vaidya [1], Gautam Mudaliar[2], Abhishek Thombare[3] and Sumit Chavan[4]
[1-4]RMD Sinhgad School of Engineering, Pune, India
Email: nimishvaidya99@gmail.com; gautammudliar@gmail.com

***Abstract*—Real Estate / Land is the least straight-forward industry in our society. House costs keep changing everyday consistently and from time to time are publicized instead of being established on valuation. House estimation process is a critical purpose of land which can be used to gain profits. The composition tries to get supportive gaining from valid data of property markets. Computer based intelligence strategies are applied to explore evident property trades to discover profitable models for house buyers and vendors. For evaluating the cost various regression techniques were attempted and one with most raised exactness is used.**

***Index Terms*— regression, ridge regression, lasso regression, random forest tree, prediction, parameters.**

I. INTRODUCTION

The objective of the project is to develop the optimum machine learning model for predicting the price of property on a dataset of house sale prices in King County, Seattle from May 2014 to May 2015. The dataset provides various features houses have like locality, condition, size, age, etc, and the prices at which they were sold.

II. RELATED WORK

The land business has turned into an aggressive and nontransparent industry. The information mining process in such an industry gives a preferred position to the engineers by preparing those information, estimating future patterns and consequently helping them to settle on ideal learning driven choices. Our primary concentration here is to build up a model which predicts the property cost for a client as indicated by his\her interests. Our model examinations a lot of parameters chosen by the client to locate a perfect value as per their necessities.
For this it uses techniques called linear regression, ridge regression, lasso regression, random forest tree, for prediction and tries to gives an analysis of the results obtained. It helps to establishes the relationship between dependent variable and other changing independent variable also known as label attribute and regular attribute respectively.

III. PROPOSED SYSTEM

Our dataset involves different basic parameters and information mining has been at the base of our framework. We at first tidied up our whole dataset and furthermore truncated the exception esteems. Further,

Grenze ID: 01.GIJET.6.2.502

we gauged every parameter dependent on its significance in deciding the estimating of the framework and this drove us to build the worth that every parameter retains in the framework. We shortlisted 5 diverse AI calculations and tried our framework with various mixes that can ensure best potentially dependability of our outcomes.

IV. VARIABLES

A total of 21 parameters were considered out of which price was dependent and rest all were independent parameters.
id (a notation for a house), date (Date house was sold), price (Price is prediction target), bedrooms (Number of Bedrooms/House), bathrooms (Number of bathrooms/House), sqft_living (square footage of the home), sqft_lot (square foot of the lot), floors (Total floors (levels) in house), waterfront (House which has a view to a waterfront), view (Has been viewed), condition (How good the condition), grade (overall grade given to the housing unit, based on King County grading system), sqft_above (square footage of house apart from basement), sqft_basement (square footage of the basement), yr_built , yr_renovated (Year when house was renovated), Zipcode, lat (Latitude coordinate), long (Longitude coordinate), sqft_living15 (Living room area in 2015(implies—some renovations) This might or might not have affected the lotsize area), sqft_lot15 (lotSize area in 2015(implies some renovations))

TABLE I. CO-RELATION

Parameter	**Price**	**Parameter**	**Price**
price	1	waterfront	0.26636943
sqft_living	0.70203505	floors	0.25679388
grade	0.66743425	yr_renovated	0.12643379
sqft_above	0.60556729	sqft_lot	0.08966086
sqft_living15	0.58537890	sqft_lot15	0.08244715
bathrooms	0.52513750	yr_built	0.05401153
view	0.39729348	zipcode	-0.0532028
sqft_basement	0.32381602	condition	0.03636178
bedrooms	0.30834959	long	0.02162624
lat	0.30700348	id	-0.0167621

This table shows co-relation values of different parameters with parameter price.

V. ALGORITHMS

A. Linear Regression

Linear regression is the most straight forward technique for forecast. It utilizes two things as factors which are the predictor variable and the other variable which is the most essential one. These regression evaluations are utilized to clarify the connection between one dependent variable and at least or more independent variables. The equation of the regression equation with one dependent and one independent variable is defined by the formula.

$$b = y + x*a \qquad (1)$$

where, b = estimated dependent variable score, y = constant, x = regression coefficient, and a = score on the independent variable.

B. Forest Regression

It uses technique called as Bagging of trees. The main idea here is to decorrelate the several trees. We then reduce the Variance in the Trees by averaging them. Using this approach, a large number of decision trees are created. Random forest training algorithm applies the technique of bootstrap aggregating, to tree learners.
Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: For b = 1, ..., B:
1. Sample, with replacement, n training examples from X, Y; call these Xb, Yb.
2. Train a classification or regression tree fb on Xb, Yb.

C. Ridge Regression

It is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. Ridge regression equation is written in matrix form as

$$Y = XB + e \quad (2)$$

where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals.

D. Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, similar to the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

VI. CONCLUSIONS

The best model performance is given by Linear Regression with .88 R-Squared value and minimum RMSE Values. Thus we can conclude that 88% of variations in dependent variable, ie price, are explained by the independent variables present in our model.

Name	rsquared	rmse
Linear Regression	0.88	127519.31
Ridge Regression	0.87	132052.04
Lasso Regression	0.85	142393.73
Decistion Tree	0.66	214681.94
Random Forest Tree	0.75	183825.07

REFERENCES

[1] J. J. WANG, “Predicting House Price With a Memristor-Based Artificial Neural Network”
[2] Ayush Varma, “House Price Prediction Using Machine Learning And Neural Networks”
[3] “Housing Price Prediction using Machine Learning Algorithms”
[4] Rushab Sawant, “A Multi Feature Based Housing Price Prediction for Indian Market Using Machine Learning”
[5] Muhammad Fahmi Mukhlishin, “Predicting House Sale Price Using Fuzzy Logic, Artificial Neural Network and K-Nearest Neighbour”
[6] Yingyu Feng, “Comparing Multilevel Modelling and Artificial Neural Networks in House Price Prediction”