International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-5, Issue-6, Jun.-2018
http://iraj.in

VALUATION OF HOUSE PRICES USING PREDICTIVE TECHNIQUES

[1]NEELAM SHINDE, [2]KIRAN GAWANDE

[1]Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India.
[2]Assistant Professor, Sardar Patel Institute of Technology, Mumbai, India

Abstract - In this paper, we are predicting the sale price of the houses using various machine learning algorithms. Housing sales price are determined by 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. This paper uses machine learning algorithms to build the prediction model for houses. Here, machine learning algorithms such as logistic regression and support vector regression, Lasso Regression technique and Decision Tree are employed to build a predictive model. We have considered housing data of 3000 properties. Logistic Regression, SVM, Lasso Regression and Decision Tree show the R-squared value of 0.98, 0.96,0.81 and 0.99 respectively. Further, we have compared these algorithms based on parameters such as MAE, MSE, RMSE and accuracy. This paper also represents significance of our approach and the methodology.

Keywords - Real Estate, Prediction Model, Linear Regression, Support Vector Machine, Decision Tree, Lasso

I. INTRODUCTION

Real Estate Property is not only the basic need of a man but today it also represents the riches and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Changes in the the real estate price can affect various household investors, bankers, policy makers and many. Investment in real estate sector seems to be an attractive choice for the investments. Thus, predicting the real estate value is an important economic index.

India ranks second in the world in number of households according to 2011 census with a number of 24.67 crore. India is also the fastest growing major economy ahead of China with former's growth rate as 7% this year and predicted to be 7.2% in the next year. According to the 2017 version of Emerging Trends in Real Estate Asia Pacific, Mumbai and Bangalore are the top-ranked cities for investment and development. These cities have supplanted Tokyo and Sydney. The house prices of 22 cities out of 26 dropped in the quarter from April to June when compared to the quarter January to March according to National Housing Bank's Residex(residential index). With the introduction of Real Estate Regulation Development Act (RERA) and Benami property Act throughout the country India, more number of investors are attracted to invest into real estate in India. The strengthening and modernizing of the Indian economy has made India as attractive Investment destination. However, past recessions show that real estate prices cannot necessarily grow. Prices of the real estate property are related to the economic conditions of the state [1]. Despite this, we are not having proper standardized ways to measure the real estate property values.

Generally the property values rise with respect to time and its appraised value need to be calculated. This appraised value is required during the sale of property or while applying for the loan and for the marketability of the property. These appraised values are determined by the professional appraisers. However, drawback of this practice is that these appraisers could be biased due to bestowed interests from buyers, sellers or mortgages. Thus, we require an automated prediction model that can help to predict the property values without any bias. This automated model can help the first time buyers and less experienced customers to understand whether the property rates are overrated or underrated.

Now, Property prices depend on various parameters in the economy and society. However, previous analyses show that house prices are strongly dependent on the size of the house and its geographical location [2], [3]. We have also considered various intrinsic parameters (such as number of bedrooms, living area and construction material) and also external parameters (such as location, proximity, upcoming projects, etc.) [4], [5]. Then we have applied these parameter values to two different machine learning algorithms. We have considered linear regression model and support vector regression model to predict the price value of the house and compared their output.

In this paper, we are predicting house price values using two models i.e. Linear regression, support vector regression, Lasso Regression and Decision Tree with their corresponding accuracy and comparing them based on various error metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), R- Squared value and Root Mean Squared Error(RMSE). In addition to this we will also discuss the significance our approach and the methodology used

II. RELATED WORK

In last two decades forecasting the property value has become an important field. Rise in the demand for

property and unpredictable behavior of economy compel researchers to find out a way that predict the real estate prices without any biases. Thus, it is a challenge for researchers to find out all the minute factors that can affect the cost of property and make a predictive model by taking into consideration all the factors. Building a predictive model for real estate price valuation requires a thorough knowledge on the subject. Many researchers have worked on this problem and communicated their research work.

Most of this research work is inspired from [6]. The author has scraped the housing data set from Centris.ca and duProprio.com. Their dataset consists of approximately 25,000 examples and 130 factors. Around 70 features were scraped from the above websites and real estate agencies such as, RE/MAX, Century 21, and Sutton, etc. Other 60 features were sociodemographic based on where the property is located. Later, author implemented Principal Component Analysis to reduce the dimensionality. The author used four regression techniques to predict the price value of the property. The four techniques are Linear Regression, Support Vector Machine, K-Nearest Neighbors (KNN) and Random Forest Regression and an ensemble approach by combining KNN and Random Forest Technique. The ensemble approach predicted the prices with least error of 0.0985. However, applying PCA did not improve the prediction error.

A lot of researches have been done on Artificial Neural Networks. This has helped many researchers focusing on real estate problem to solve using neural networks. In [7], the author has compared hedonic price model and ANN model that predict the house prices. Hedonic price models are basically used to calculate the price of any commodity that are dependent on internal characteristics as well as external characteristics. The hedonic model basically involves regression technique that considers various parameters such as area of the property, age, number of bedrooms and so on. The Neural Network is trained initially and the weights and biases of the edges and nodes respectively are considered using trial and error method. Training the Neural Network model is a black box method. However, the R-Squared value for Neural Network model was greater compared to hedonic model and the RMSE value of Neural Network model was relatively lower. Hence it is concluded that Artificial Neural Network performs superior than Hedonic model.

Some researchers like that in [8] have used classifiers to predict the property values. The author in research article [8] has collected the data from Multiple Listing Service (MLS), historical mortgages rates and public school ratings. Real Estate Data was obtained from Metropolitan Regional Information Systems (MRIS) database. The author extracted approximately 15,000 records from these three sources which included 76 variables. Subsequently, t-test was used to select 49 variables as a preliminary screening. Their research question was to determine whether the closing price was higher or lower than the listing price [8]. Thus to address this classification problem, the author used four machine learning models. C4.5, RIPPER, Naive Bayesian, and AdaBoost are the four algorithms used by author. However, they found that RIPPER outperforms other house prediction models. However the drawback is that performance evaluation is based only on classifiers. Performance comparison of other machine learning algorithms should also be considered.

In article [9], the authors have predicted the stock market prices using linear regression technique. They have collected stock market data from TCS stock Database. The author have also used RBF and polynomial regression technique along with linear regression and found that latter is superior to the remaining techniques.

In [11], the author has considered the most macroeconomic parameters that affect the house prices variation. In this, the author has used back propagation neural network (BPN) and radial basis function neural network (RBF) to establish the nonlinear model for real estates price variation prediction. The dataset have been taken from Taipei, Taiwan based on leading and simultaneous economic indices. The author has considered 11 parameters. The prediction results obtained from them are compared to public Cathay House Price Index or the Sinyi Home Price Index. The two error metrics used were Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). When the prediction results were compared to Cathay House Price Index, RBF Neural Network showed better prediction results than BPN Neural Network. Similarly, for Sinyi Home Price Index BPN Neural Network showed better prediction results than RBF Neural Network.

Some research articles describe the in depth methods and procedures to collect the real estate data and their pre-processing techniques. The author in article [12] describes software that is used in real estate price evaluation. The software analyzes various real estate servers and web pages of real estate companies, and records their current links to real estate purchase or rental into their software database. He has gathered data from Czech Republic. The data is gathered every month to record the changes happening in real estate. The software gathers 110,000 entries every month. These entries include various texts, advertisements and images of the property. The author has collected data from the year 2007 to 2015.This unstructured data that is collected is exported into a structured form. Different property types have different parameters. Thus it makes the data set more readable. This data set is then evaluated. New entries made each month are compared to the older entries and checked for their completeness. In the final phase of the software, this available clean data set is then evaluated and produces various visualizations according to the requirement of the user. Thus the

output obtained may be used as grounds for appropriate investments and/or housing decisions for both common persons and companies.

Some researchers have focused on feature selection and feature extraction procedure. The author in article [14] uses a open source data set of the housing sales in King County, USA. There are about 20 explanatory variables. The author has compared various feature selection and feature extraction algorithms combined with Support Vector Regression. The author has collected approximately 21,000 observations in a time period of one year. The paper shows various data analysis performed on the data set. Feature Selection is the process of selecting a subset of variables from a given set of parameters either based on their importance or their frequency. However, feature extraction is the process of reducing the dimensionality of the data. Initial set of data is transformed into derived values which are equally informative and non-redundant. The three feature selection algorithms used are Recursive Feature Elimination (RFE), Lasso and Ridge and Random Forest Selector and the mean from each algorithm is calculated. Using feature selection, the author selects fifteen features out of twenty. The feature extraction algorithm used is Principal Component Analysis (PCA) and reduces the parameters from twenty to sixteen. However, the author found that both the techniques work equally well with the R squared value of 0.86.

III. METHODOLOGY

Methodology represents a description about the framework that is undertaken. It consists of various milestones that need to be achieved in order to fulfill the objective. We have undertaken different data mining and machine learning concepts. The following diagram (figure 1) represents step-wise tasks that need to be completed:

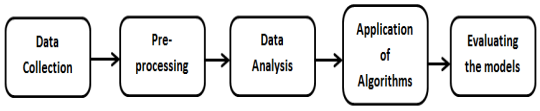

Figure 1: System Architecture

1) Data Collection

The dataset used in this project was an open source dataset from KaggleInc [22]. It consists of 3000 records with 80 parameters that have the possibility of affecting the property prices. However out of these 80 parameters only 37 were chosen which are bound to affect the housing prices. Parameters such as Area in square meters, Overall quality which rates the overall condition and finishing of the house, Location, Year in which house was built, Numbers of Bedrooms and bathrooms, Garage area and number of cars that can fit in garage, swimming pool area, selling year of the house and Price at which house is sold. Selling price is a dependent variable on several other independent variables. Some parameters had numerical values and some were ratings. These ratings were converted to numerical values. Following Table 1 represent a brief description about most important parameters that affect the selling price of the house.

Parameters	Description	Datatype
OverallQual	Rates the overall material and finish of the house	Numerical
YearBuilt	Original construction date	Numerical
TotalBsmtSF	Total square feet of basement area	Numerical
GrLivArea	Above grade (ground) living area square feet	Numerical
FullBath	Full bathrooms above grade	Numerical
GarageCars	Size of garage in car capacity	Numerical
GarageArea	Size of garage in square feet	Numerical
WoodDeckSF	Wood deck area in square feet	Numerical
PoolArea	Pool area in square feet	Numerical
YrSold	Year Sold (YYYY)	Numerical
SalePrice (Dependent Variable)	Selling Price of the house	Numerical

Table 1
The Parameters

2) Data Preprocessing

It is a process of transforming the raw, complex data into systematic understandable knowledge. It involves the process of finding out missing and redundant data in the dataset. Entire dataset is checked for NaN and whichever observation consists of NaN will be deleted. Thus, this brings uniformity in the dataset. However in our dataset, there was no missing values found meaning that every record was constituted its corresponding feature values.

3) Data Analysis

Before applying any model to our dataset, we need to find out characteristics of our dataset. Thus, we need to analyze our dataset and study the different parameters and relationship between these parameters. We can also find out the outliers present in our dataset. Outliers occur due to some kind of

experimental errors and they need to be excluded from the dataset.

From the analysis we found out that there existsone or two outliers. The general trend for Sale price over different parameters. 'GrLivArea' and 'TotalBsmtSF' seem to be linearly related with 'SalePrice'. The overall quality of the house and Area rises the sale price of the house rises too! However, Overall quality and number of bathrooms are non-correlated and are independent of each other. Total Basement Area and Ground Living Area are correlated to each other. There exists an outlier in all the graphs of Total Basement Area. This outlier could be present due to experimental errors and hence that observation can be avoided.

A correlation number gives the degree of association between two variables. The correlation number exists between +1 to -1. A positive number represents a positive correlation between two variables and vice versa. However, if the correlation number is 0, it shows that there is no correlation between two variables and they are independent of each other. Correlation Matrix gives a in depth idea about correlation among various parameters. We have plotted a correlation matrix for 37 selected parameters in the following figure 3.

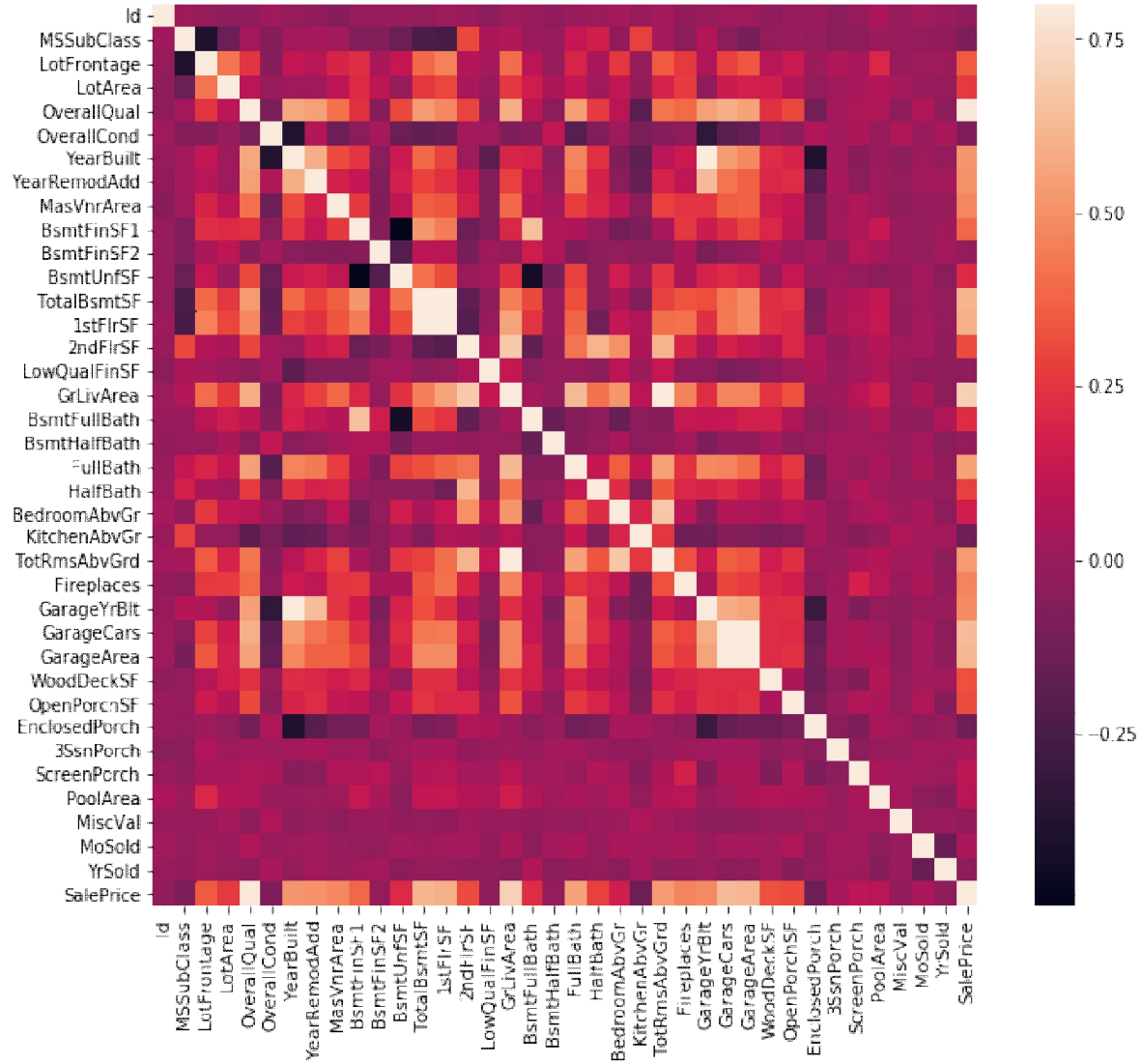

Figure 3: Correlation Matrix

Blocks with correlation number towards zero show a higher variance since the correlation is 0 and are independent. These parameters cannot be neglected. However, numbers towards +1 or -1 are showing higher relation between variable thus either of them can be neglected to minimize the number of parameters.

4) Application of Algorithms

Once the data is clean and we have gained insights about the dataset, we can apply an appropriate machine learning model that fits our dataset. We have selected four algorithms to predict the dependent variable in our dataset. The algorithms that we have selected are basically used as classifiers but we are

training them to predict the continuous values. The four algorithms are Logistic Regression, Support Vector Machine, Lasso Regression Technique and Decision Tree. These algorithms were implemented with the help of python's SciKit-learn Library [15]. The predicted outputs obtained from these algorithms were saved in comma separated value file. This file was generated by the code at run time.

1. Logistic Regression

To establish baseline performance with a logistic classifier, we used Logistic Regression to predict the price targets, Si , as a logistic function of the data, X

$$S_i = \frac{1}{1+e^{-X_i\beta+\varepsilon_i}} \quad (1)$$

S_i represents the continuous variable which is bordered between zero and one. Xi represents the independent data and represents some error which is symmetrically distributed around zero and variance, 2 Our dependent variable is continuous in nature which is not bounded. However, this function gives us a binary output whose probability is bounded between zero and one, thus we transform this logistic distribution into a simple linear regression function [20].

$$\log\left(\frac{S_i}{1-S_i}\right) = X_i\beta + \varepsilon_i \quad (2)$$

Let us assume that S_i is a logistic function of a vector of independent variables, X. The actual values of Si can then be described by the equation:

$$S_i = \frac{1}{1+e^{-X_i\beta}} \quad (3)$$

But, now let us consider that instead of noting the actual values Si, the variable Si^* is observed where

$$S_i^* = S_i + \varepsilon_i \quad (4)$$

Constructing the logit regression with the dependent variable S* gives the regression equation:

$$\log\left(\frac{S_i^*}{1-S_i^*}\right) = X_i\beta + u_i \quad (5)$$

Thus the resulting error term can be given by:

$$u_i = \log\left(\frac{S_i^*}{1-S_i^*}\right) - \log\left(\frac{S_i}{1-S_i}\right) \quad (6)$$

The predicted prices by Logistic Regression are given in the figure 4.

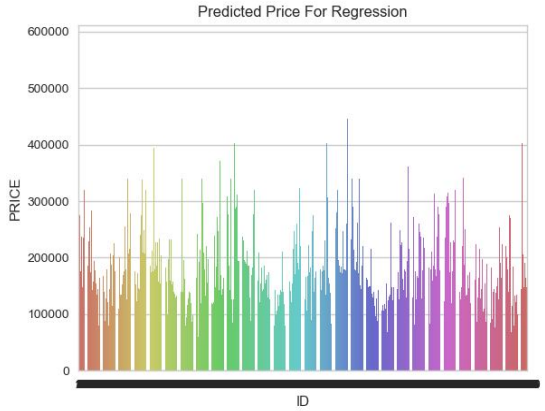

Figure 4: Predicted price by Logistic Regression

2. Support Vector Regression

Support Vector Machine can also be used as a regression technique. It uses the same principles as that of support vector machine classifier. However, in Linear Support Vector Regression it estimates a function that maximizes the deviation from the actual target Y_n within normalized margin strip, , by keeping the function as flat as possible. Thus, Support Vector Regression Technique is a convex minimization problem that finds the normal vector $w \in R^M$ of the linear function as follows [6]:

$$\text{minimize}\left(\frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\gamma_n + \gamma_n^*\right)$$

(7)

Subject to the constraint for each n:

$$y_n - (w * X_n) \leq \epsilon + \gamma_n \quad (8)$$

$$(w * X_n) - y_n \leq \epsilon + \gamma_n^* \quad (9)$$

$$\gamma_n, \gamma_n^* \geq 0 \quad (10)$$

γ,γ_n represent the 'slack' variables. The constant C > 0 determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated [17]. Following figure represents predicted house prices by Support Vector Regression.

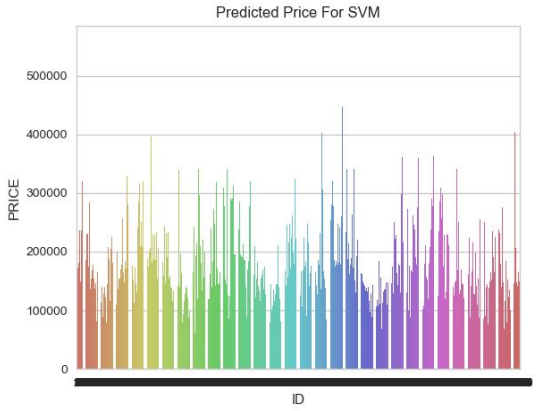

Figure 5: Predicted price by Support Vector Regression

3. Lasso Regression

Lasso is a powerful regression technique. It works by penalizing the magnitude of coefficients of features along with minimizing the error between predicted and actual observations. Lasso is called as L1 Regularization technique. Lasso attempts to minimize the cost function. The cost function is given as

Cost(W)= RSS(W) + **α** (Sum of squares of weight)

Here RSS refers to 'Residual Sum of Squares' meaning the sum of square of errors between the predicted and actual values in the training data set. **α** is co-efficient that takes various values. There are three cases for values of **α**.

1. **α** = 0; Same coefficients as simple linear regression
2. **α** = ∞; All coefficients zero

3. $0 < \alpha < \infty$: coefficients between 0 and that of simple linear regression

The Lasso function can be mathematically be given as follows:

$$\text{cost}(w) = \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j x_{ij} \right\}^2 + \alpha \sum_{j=0}^{M} |w_j| \quad (11)$$

The Lasso function implemented predicts the target parameter as follows:

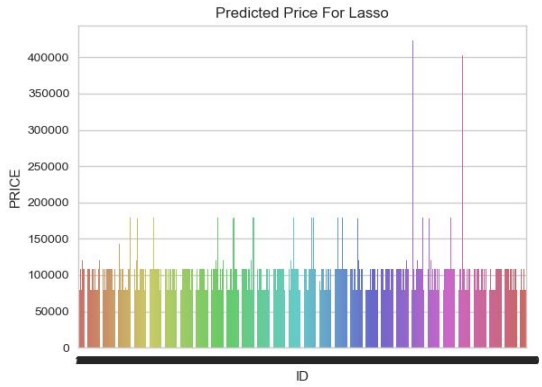

Figure 6: Predicted price by Lasso Regression

4. Decision Tree

Decision trees are considered to be the best and most widely used supervised learning algorithm. This model has the ability to predict the output with at most accuracy and stability. It is used to predict any kind of problems such as classification or regression. However, in our case we want to predict a continuous target value hence our problem is of regression type.

In this model, the available dataset can be continuous or categorical. We use binary tree that will recursively partition the predictor vector into different subsets such that our target value **y** is more homogenous. $\boldsymbol{x}$ represents the vector of predictors $x=x_1,x_2,x_3,...,x_n$.A decision tree with **t** terminal nodes is used for communicating the classification decision. A parameter$\varphi=(\varphi_1,\varphi_2,\varphi_3,\ldots.\varphi_t)$associates the parameter value $\varphi_i(i=1,2,3,...,t)$with the i^{th} terminal node. The partitioning procedure searches through all values of predictor variables (vector of predictors) to find the variable x that provides best partition into child nodes [13]. The best partition will be the one that minimizes the weighted variance.

However one of the key challenges in decision trees is overfitting. In the worst case, it will consider leaf node for each value and thus give 100% accuracy. In order to prevent overfitting we can set constraints on the size of the tree or pruning the tree. The following graph represents values predicted by decision tree for our dataset:

Figure 7: Predicted price by Decision Tree

RESULTS

The following section shows the results of various algorithms that are applied. We have taken into consideration different performance metrics such as Accuracy, R-squared value, Root Mean Squared Value (RMSE), Mean Absolute Value (MAE) and Mean Squared Value (MSE). Using these parameters we have compared the four models in Table 2.

	Accuracy	R-Square	RMSE	MAE	MSE
LR	72.81%	0.987	8922	6118	79604145
SVR	67.81%	0.968	14101	76429	1.99E+08
Lasso	60.32%	0.81	34275	21058	1.7E+08
DT	84.64%	0.99	217	5.68	47184.93

Table 2 Results

From the above table we find that Decision Tree gives a higher accuracy and R-squared value and low error values.

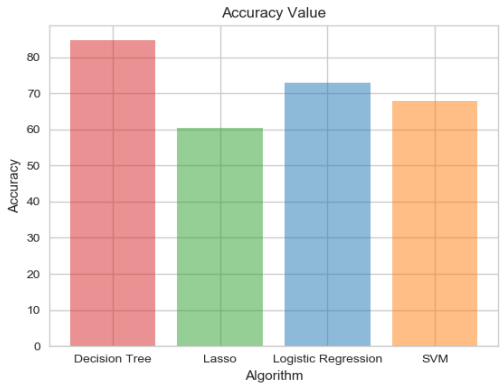

Figure 8: Accuracy Values

On comparing the various models, we find that decision tree works the best with highest accuracy of 84.64% and Lasso performs least with an accuracy of 60.32%.

Decision Tree produces hardly any error with RMSE value of 217 and Lasso performs worst with its RMSE value as 34245.

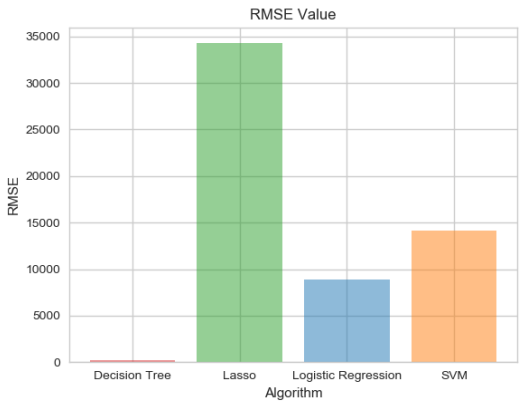

Figure 9: RMSE Values

Thus, we can conclude that Decision Tree overfits our dataset and gives a very high accuracy but this can be a problem because it is also taking into consideration the various noises present around. Also we have divided our dataset in the ratio 50:50 to avoid the problem of overfitting from our side.

CONCLUSION

In this research paper, we have used machine learning algorithms to predict the house prices. We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the parameters. Thus we can select the parameters which are not correlated to each other and are independent in nature. These feature set were then given as an input to four algorithms and a csv file was generated consisting of predicted house prices. Hence we calculated the performance of each model using different performance metrics and compared them based on these metrics. We found that Decision Tree overfits our dataset and gives the highest accuracy of 84.64%. Lasso gives the least accuracy of 60.32%. Logistic Regression and Support Vector Regression giving an accuracy of 72.81% and 67.81% respectively Thus we conclude that we implemented classifiers to the problem of regression to check how well can classifier fit to regression problem [21].

For future work, we recommend that working on large dataset would yield a better and real picture about the model. We have undertaken only few Machine Learning algorithms that are actually classifiers but we need to train many other classifiers and understand their predicting behavior for continuous values too. By improving the error values this research work can be useful for development of applications for various respective cities.

REFERENCES

[1] R. J. Shiller, “Understanding recent trends in house prices and home ownership,” National Bureau of Economic Research, Working Paper 13553, Oct. 2007. DOI: 10.3386/w13553. [Online]. Available: http://www.nber.org/papers/w13553.

[2] D. Belsley, E. Kuh, and R. Welsch, Regression Diagnostics: Identifying Influential Data and Source of Collinearity. New York: John Wiley, 1980.

[3] J. R. Quinlan, “Combining instance-based and model-based learning,” Morgan Kaufmann, 1993, pp. 236–243.

[4] S. C. Bourassa, E. Cantoni, and M. Hoesli, “Predicting house prices with spatial dependence: a comparison of alternative methods,” Journal of Real Estate Research, vol. 32, no. 2, pp. 139–160, 2010. [Online] Available:http://EconPapers.repec.org/RePEc:jre:issued:v:32:n:2:2010:p:139-160.

[5] S. C. Bourassa, E. Cantoni, and M. E. Hoesli, “Spatial dependence, housing submarkets and house price prediction,” eng, 330; 332/658, 2007, ID: unige:5737.[Online]. Available: http:// archive - ouverte. unige. ch/unige:5737.

[6] Pow, Nissan, Emil Janulewicz, and L. Liu. "Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal." (2014).

[7] Limsombunchai, Visit. "House price prediction: hedonic price model vs. artificial neural network."New Zealand Agricultural and Resource Economics Society Conference. 2004.

[8] Park, Byeonghwa, and Jae Kwon Bae. "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data."Expert Systems with Applications 42.6 (2015): 2928-2934.

[9] Bhuriya, Dinesh, et al. "Stock market predication using a linear regression." Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of.Vol. 2.IEEE, 2017.

[10] Majumder, Manna, and MD Anwar Hussian. "Forecasting of Indian stock market index using artificial neural network."Information Science (2007): 98-105.

[11] Li, Li, and Kai-Hsuan Chu. "Prediction of real estate price variation based on economic parameters." Applied System Innovation (ICASI), 2017 International Conference on.IEEE, 2017.

[12] Hromada, Eduard. "Mapping of real estate prices using data mining techniques." Procedia Engineering 123 (2015): 233-240.

[13] Razi, Muhammad A., and KuriakoseAthappilly. "A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models." Expert Systems with Applications 29.1 (2005): 65-74.

[14] Wu, Jiao Yang. "Housing Price prediction Using Support Vector Regression." (2017).

[15] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011): 2825-2830.

[16] Manning, Richard L. "Logit regressions with continuous dependent variables measured with error." Applied Economics Letters 3.3 (1996): 183-184.

[17] Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." Statistics and computing 14.3 (2004): 199-222.

[18] Jaen, Ruben D. "Data Mining: An Empirical Application in Real Estate Valuation." FLAIRS Conference. 2002.

[19] Lim, Wan Teng, et al. "Housing price prediction using neural networks." Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on.IEEE, 2016.

[20] Manning, Richard L. "Logit regressions with continuous dependent variables measured with error." Applied Economics Letters 3.3 (1996): 183-184.

[21] Torgo, Luis, and Joao Gama. "Regression using classification algorithms." Intelligent Data Analysis 1.4 (1997): 275-292.

[22] https://www.kaggle.com/ohmets/feature-selection-for-regression/data

★★★