HOUSING PRICE PREDICTION

Minor Project I

Submitted by:

Saurabh Kumar Gupta (9917103236)
Aastha Juneja (9917103239)
Ashutosh Baghel (9917103255)
Himankur Goyal (9917103266)

Under the supervision of:
Dr. Arti Jain

Department of CSE/IT
Jaypee Institute of Information Technology University , Noida

NOVEMBER 2019

Abstract

Housing price prediction is computerized prediction model which can help to forecast the property values precisely without any bias and also elucidates of whether the property rates are overrated or underrated. House prices increase every year,so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the investors to arrange the right time to purchase a house.

Our aim is to provide an efficient regression model that is well developed to accurately estimate the price of the house given the features so that investors will be save from being cheated. This will also help us to maintain the balance in economy in the real-estate business. The result from this research came up with the solution that which algorithm predict the price with minimum prediction error.

ACKNOWLEDGEMENT

We would like to place on record our deep sense of gratitude to Dr. Arti Jain (Ph.D.-CSE), Jaypee Institute of Information Technology, India for his/her generous guidance, help and useful suggestions.

We express our sincere gratitude to Computer science & Engineering Dept. of Jaypee Institute of Information Technology , India, for his/her stimulating guidance, continuous encouragement and supervision throughout the course of present work.

We also wish to extend our thanks to the seniors and our batchmates for their insightful comments and constructive suggestions to improve the quality of this project work.

Signature(s) of Students

Himankur Goyal (9917103266)

Aastha Juneja (9917103239)

Saurabh Kumar Gupta (9917103236)

Ashutosh Baghel (9917103255)

Table of contents

List of Figures

Nomenclature

Symbol

Y	Dependent variable(Criterion variable which we ar predicting)
X	Independent variable (Predictor variable on which we base our predictions)
a	Intercept
b	Slope of a line
S(T, X)	Standard deviation for two attributes(target and predictor)
SDR(T, X)	Standard deviation reduction for two attributes(target and predictor)

1.INTRODUCTION

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project. house prices will be predicted given explanatory variables that cover many aspects of residential houses.

1.1 Problem Statement

There is a problematic issue when the evaluators determine the assessed values. These employed people could be partial due to the extra interest they gain from the buyers and sellers. Despite this, there is no standardized ways to measure the housing prices. Thus, it becomes important to develop a computerized prediction model which can help to forecast the property values precisely without any bias and also elucidates of whether the property rates are overrated or underrated. Thus, predicting the real estate value is an important economic index.

1.2 Purpose & Objective

Buyers would use house price prediction to search for candidate houses that match their financial capabilities. Similarly, house owners would need it to keep monitoring the market and seek the best opportunity for house selling. Moreover, real estate sales agents also rely on house price prediction to help customers better find out market trends, and the accuracy of prediction has become an important criterion for measuring the credibility of house sales agents. Usually there is increase in the value of the property with respect to time and its assessed value need to be estimated. There is a requirement of this calculated value during the sale of property or while applying for the loan. Predicting the real estate value is an important economic index.

So to make our contribution in this field we build this project model which predict the future values of houses on the basis of past scenarios and investor's requirements.

Growth in the demand for property and irregular behavior of economy oblige us to work on the housing prediction models. Basically, finding out the minute factors which can affect the cost of property and making the respective predictive model is a need of an hour so that government can make a reasonable urban planning. Our objective is to create a regression model that is well developed to accurately estimate the price of the house given the features so that investors will be save from being cheated. This will also help us all to maintain the balance in economy in the real-estate business.

1.3 Summary

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 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. we have considered some intrinsic factors such as number of bedrooms, bathrooms, floors and so on… and also the extrinsic factors such as the latitude and longitude (basically the location), year built, etc. Modeling is usually carried out from a global view, i.e. implemented on the entire house data directly. .We have applied these features to various machine learning algorithm in order to predict the prices. As continuous house prices, they will be predicted with various regression techniques including linear regression with gradient boosting, decision tree and Random Forest Tree regressor.These models will predict the value of the house. On the basis of their accuracy, the models will be compared and choosen for predicting the real estate prices.

2.BACKGROUND STUDY

2.1 Literature Review

The related works are on the house price prediction as follows:

Nissan Pow, Emil Janulewicz, Liu (Dave) Liu, 2016, “Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal “:They predicted both asking and sold prices of real estate properties based on features such as geographical location, living area, and number of rooms, etc. Additional geographical features such as the nearest police station and fire station were extracted from the Montreal Open Data Portal, the final price sold was also predicted with an error of 0.023 using the Random Forest Regression[1].

Yu, Jiafu Wu. 2016, “Real Estate Price Prediction with Regression and Classification”: Yu, Jiafu Wu. has predicted house prices given explanatory variables that cover many aspects of residential houses. As continuous house prices, they are predicted with various regression techniques including Lasso, Ridge, SVM regression, and Random Forest regression; as individual price ranges, they are predicted with classification methods including Naive Bayes, logistic regression, SVM classification, and Random Forest classification[3].

Worzala et al. (1995) take on a contrary position and cast some doubt upon the role of neural networks compared to the traditional regression models: Worzala et al. (1995), on the other hand, take on a contrary position and cast some doubt upon the role of neural networks compared to the traditional regression models. The authors argued that even when the same data is used, results from models prepared by different neural network software package could be inconsistent and did not always outperform regression models. Lenk et al. (1997) also reached the similar conclusions. Their study documented very similar performance between the hedonic model and the neural network models[10].

Park & Bae, 2015.Analyze different ML algorithms such as RIPPER, Naïve Bayes, AdaBoost to improve prediction model.: Park’s paper analyzes the housing data on 5359 townhouses in Firfax County, Virgina based on different machine learning algorithms such as RIPPER (Repeated Incremental Pruning to Produce Error Reduction), Naïve Bayes, AdaBoost. He proposes an improved prediction model to help sellers make their decisions on the house price valuations. He concludes RIPPER algorithm outperforms other models on predicting house price (Park & Bae, 2015)[8].

3.REQUIREMENT ANALYSIS

Requirement analysis is the process of defining the expectation of the user for an application that is to be build or to be modified. A functional requirement describes what a software system should do, while non-functional requirements place constraints on how the system will do so.

3.1 Functional requirement

3.1.1 Predicted price shown by the graphs.

3.1.2 Predict the price accuracy using the different algorithm.

3.1.3 Use of JUPYTER compiler.

3.2 Non-Functional requirement

3.2.1 Predict the price accurately & efficiently.

3.2.2 Predict the price in lesser time.

3.2.3 Use of PYTHON language.

4. DETAILED DESIGN

4.1 Dataset

We have take this dataset from America household company containing around 21 columns and 21613 rows, columns are: id, date, price, bedrooms, bathroom, sqft_living,sqft_lot, floors, waterfront, view, zipcode, condition, grade, yr_built, yr_renovated, lat, long, sqft_living15, sqft_lot15, sqft_above, sqft_basement. We applied are algorithms to this dataset.

4.2 Dataset Refining & Cleaning

We learn about some columns which are of no use to our project and are the reason of outliers ,columns are waterfront, view, zipcode, yr_renovated, sqft_lot15.

IRRELEVANT AND REDUNDANT FEATURES TO BE IGNORED IN THE FUTURE ANALYSIS:

i)Waterfront

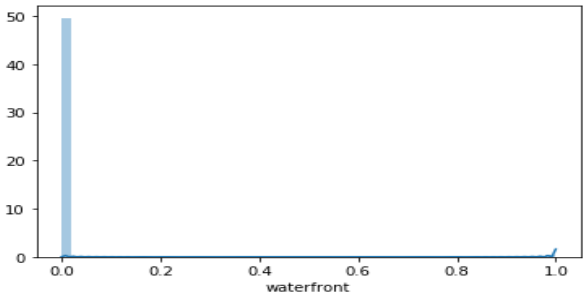

Fig.4.2.1 Distribution graph of waterfront

As it is clearly visible from above distribution graph of waterfront that maximum houses have no waterfront hence we discard this attribute from our dataset.**ii)View**

Similar case is with the view so we also discard this element from our dataset.

iii) Year Renovated

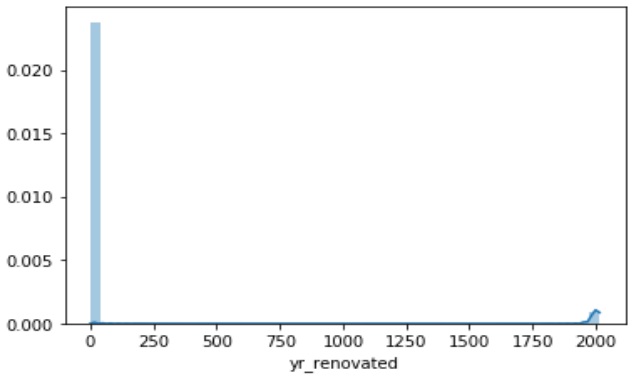

Fig.4.2.2 Distribution graph of year renovated

From above distribution graph we are sure that there is no use of yr renovated so we also discard this feature from our dataset.

iv) Zipcode

This feature is categotical,indicating the area the property sold is in. Whilst the area of a property is a key determinant in its value,this information is already recorded in Latitude and Longitude. So this is a redundant feature as a result we discard this feature also.

v) SQFT_LOT15

Owing to the questionable values of the feature due to some reasons,this feature will be ignored in any further analysis.

MOST IMPORTANT & DEPENDENT FEATURES IN DATASET:

Here we plot the Heatmap for our dataset to find out the most important feature in our dataset:

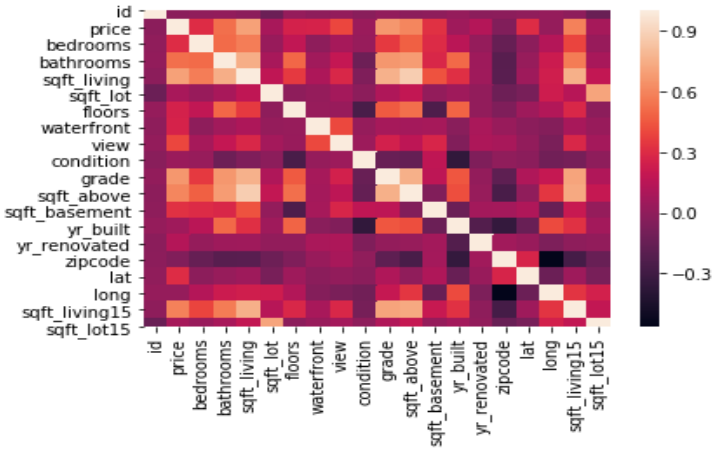

Fig.4.2.3 Heat map of correlation of variables in dataset

We can see from the heatmap that the most important features are the size of the living area(sqft_living), the grade and the year of construction,with older houses being worth less than newer builds.

1	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
4975	3438502715	20140730T000000	385000	4	3	2090	5102	1
4976	280610020	20140902T000000	825000	4	3.25	4110	14219	2
4977	2487200938	20141126T000000	815000	5	3.25	3230	5000	2
4978	1775500362	20141013T000000	625000	4	2.5	2601	34335	2
4979	2483200010	20141007T000000	690000	3	1.75	2070	6000	1
4980	4365700130	20150325T000000	210000	3	1	1660	7440	1
4981	9136101271	20150416T000000	599000	4	1	1590	4280	1.5
4982	1370804295	20150212T000000	860000	3	1.75	1860	5584	1
4983	422049178	20150212T000000	147200	3	1	1420	9600	1
4984	9161100795	20150506T000000	476900	3	1	1240	5758	1.5
4985	5561300380	20140807T000000	450000	4	2.5	2500	36254	1
4986	1443550020	20150506T000000	570000	4	2.5	2640	11816	2
4987	9547201155	20141016T000000	567500	3	1	1440	3060	1.5
4988	3625059120	20141023T000000	790000	5	3.25	3030	20446	2
4989	644000185	20140707T000000	875000	3	1.5	1820	12686	1
4990	5710500010	20140610T000000	490000	3	2	2220	10275	2
4991	2482410130	20140610T000000	335000	3	1.75	2430	9133	1
4992	6071900130	20150415T000000	550000	3	1.75	1670	10798	1
4993	6817800630	20140516T000000	385000	3	1.75	1180	10541	1
4994	6204400130	20140718T000000	395000	3	1.75	1620	8085	1
4995	9274203190	20140611T000000	650000	2	1	1030	5750	1

condition	grade	sqft_above	sqft_basement	yr_built	lat	long	sqft_living1
3	7	1350	740	1994	47.5427	-122.356	2090
4	10	2570	1540	1979	47.7382	-122.264	2760
3	9	2350	880	2002	47.5202	-122.393	1520
3	9	2601	0	1995	47.742	-122.087	2080
3	8	1340	730	1955	47.5226	-122.382	2200
3	7	1270	390	1957	47.5242	-122.362	1540
3	7	1590	0	1924	47.667	-122.335	2230
3	8	1310	550	1951	47.637	-122.4	1630
4	6	1420	0	1954	47.4232	-122.292	1400
4	6	960	280	1910	47.5675	-122.396	1460
4	8	1590	910	1978	47.4685	-122.004	2360
3	8	2640	0	1999	47.733	-121.968	2400
4	7	1440	0	1910	47.6769	-122.307	1440
3	9	2130	900	1976	47.6133	-122.106	2890
4	7	1820	0	1952	47.5886	-122.195	3020
3	9	1640	580	1980	47.5304	-122.055	2300
4	7	1410	1020	1978	47.5116	-122.157	1980
4	8	1670	0	1962	47.549	-122.17	2290
4	7	940	240	1981	47.6348	-122.032	1230
3	7	1210	410	1976	47.7349	-122.197	1700
5	8	1030	0	1928	47.5861	-122.391	1570

Fig.4.2.4 Refined Dataset

4.3 ALGORITHMS

4.3.1 Linear Regresssion

Linear regression is a statistical approach for modeling relationship between a dependent variable with a given set of independent variables. In easy words, through this model in statistics we predicts our future outcome based upon past relationship of variables. Regression works on the line equation, y= mx + c, trend line is set through the data points to predict the outcome. The variable we are predicting is called the criterion variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X.The Equation of Linear Regression:

$$Y = a + bX$$

4.3.2 Decision-tree Regressor

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Here, continuous values are predicted with the help of a decision tree regression model. 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.

$$S(T,X) = \sum P(c)S(c) \text{ where } c \in X$$

$$SDR(T,X) = S(T) - S(T,X)$$

4.3.3 Random-forest Regressor

Decision trees are computationly expensive to train carry a big risk of overfitting , and tend to find local optima because they can't go back after they have made a split.to address these weaknesses, we here use to random forest, which illustrates the power of combining many decision trees into one model. Random forest is supervised learning algorithm which uses ensemble learning method for classification.

Ensemble learning is the technique that combine the prediction from multiple machine learning algorithm together to make accurate prediction than any individual model. Model comprised of many model is called an ensemble model.

5. IMPLEMENTATION

In this part, we describe the details of technique used in the implementation of our project and describe how we apply different algorithms like Linear regression, Decision tree to train our dataset:

5.1 Linear Regression

We import our dependencies, for linear regression we use sklearn (built in python library) and import linear regression from it. We then initialize Linear Regression to a variable linreg. We again import another dependency to split our data into train and test.we've made my train data as 80% and 20% of the data to be my test data, and randomized the splitting of data by using random state. So now, we have train data, test data and labels for both let us fit our train and test data into linear regression model. After fitting our data to the model we can check the score of our data ie , prediction. in this case the prediction is **60%**

For building our prediction model, we use gradient boosting regression.

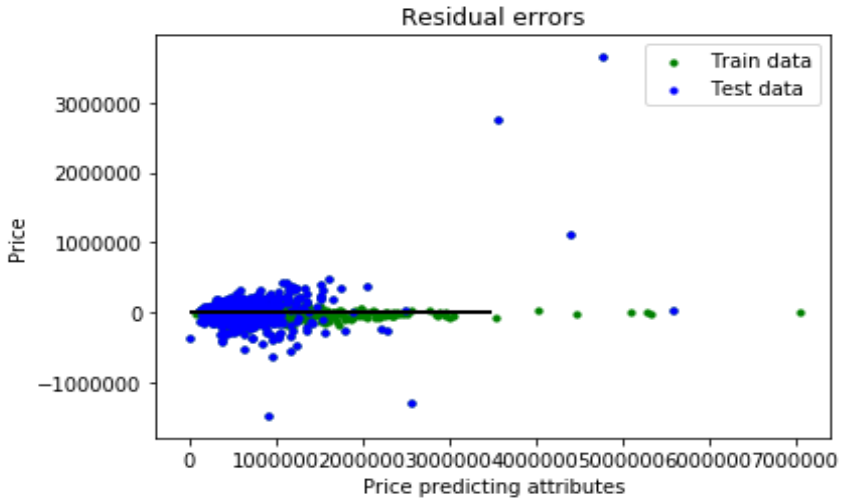

Fig.5.1 Analyzation of Linear Regressor

5.2 Decision-tree Regressor

Implementing procedure is same as we applied in above algorithm. After fitting our data to the model we can check the score of our data ie prediction. in this case the prediction is **72%.**

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(fig.5.2) represents values predicted by decision tree for our dataset:

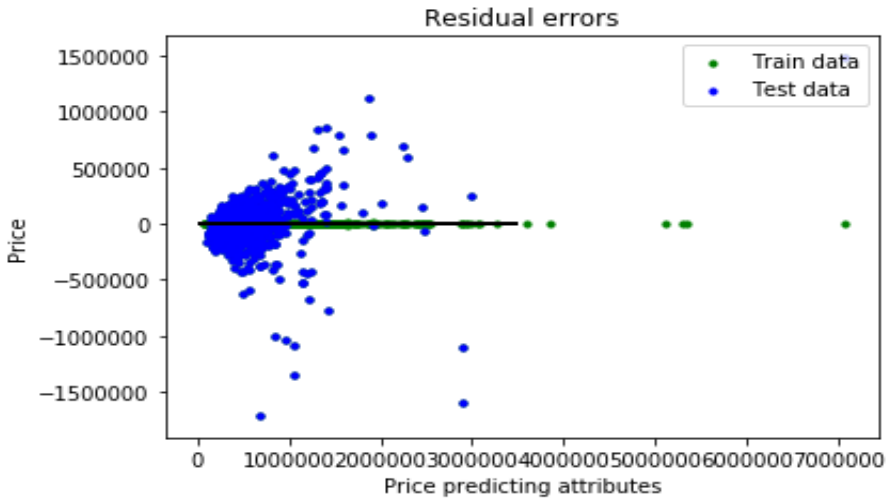

Fig.5.2 Analyzation of Decision Tree Regressor

5.3 Random-forest Regressor

General implementing procedure and splitting of training and testing data is same as in the above two algorithms. Through this algorithms we get much better results with accuracy score is **96%.** The following graph(fig.5.3) represents the values predicted by Random forest regressor:

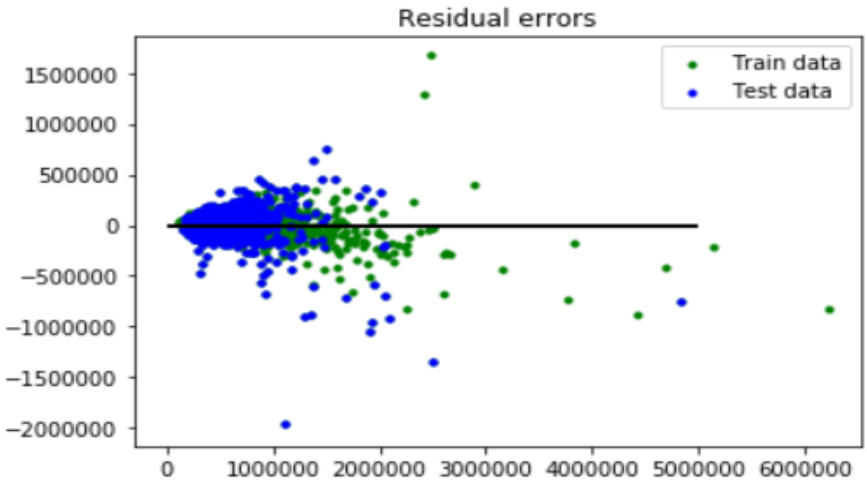

Fig.5.3 Analyzation of Random Forest Regressor

6. EXPERIMENTAL RESULTS AND ANALYSIS

After the implementation we proceed to achieve the final results i.e., the house price on the desired requirements of customer from the different algorithms. For the visualization of result with graph we take sqft_living as our main attribute for example because it's the most dependent variable in our dataset

6.1 Linear Regression

In linear regression we are successful to predict the price with an accuracy of approx. **60%.** Graphical representation of result on the basis of sqft_living:

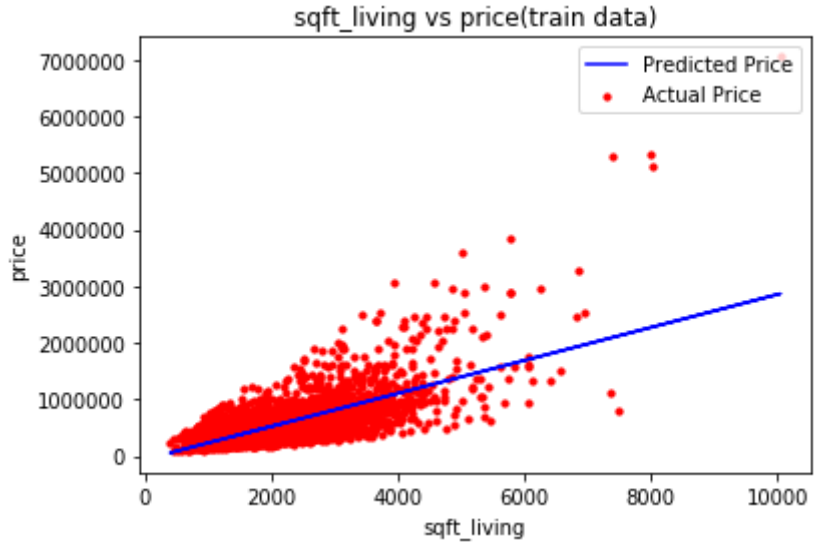

Fig.6.1.1 Price prediction on sqft_living by Linear regression(Train data)

Fig.6.1.2 Price prediction on sqft_living by Linear regression(Test data)

Similarly, an example on taking all the attributes from dataset to predict the price:On passing the values(7,3,2,1000,2000,2,3,7,1000,1000,2014,47.5112,-122.257,1340) to

the variables (id, bedrooms, bathrooms, sqft_living, sqft_lot, floors, condition, grade, sqft_above, sqft_basement, yr_built, lat, long, sqft_living15) respectively , we get the output 116227.884011 which means if the customer have the requirements as we provide in our input then the cost of his desired house will be around Rs 116227.88 lac.
In linear regression , **RMSE value** is approx. **226443.48**

6.2 Decision-tree Regressor

In Decision-tree regressor, we are successful to predict the price with an accuracy of **72%.**

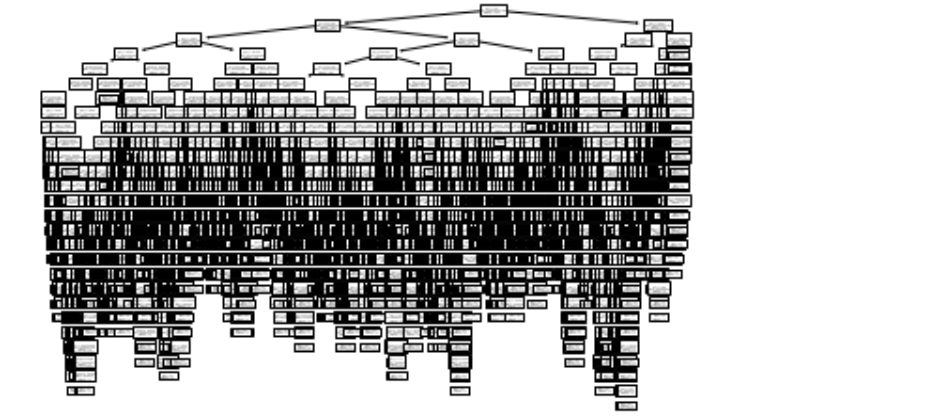

Fig.6.2.1 Tree implementation of price prediction using multi attribute by decision tree

For the clear visualization of decision tree we opt sgft_living & 10 rows

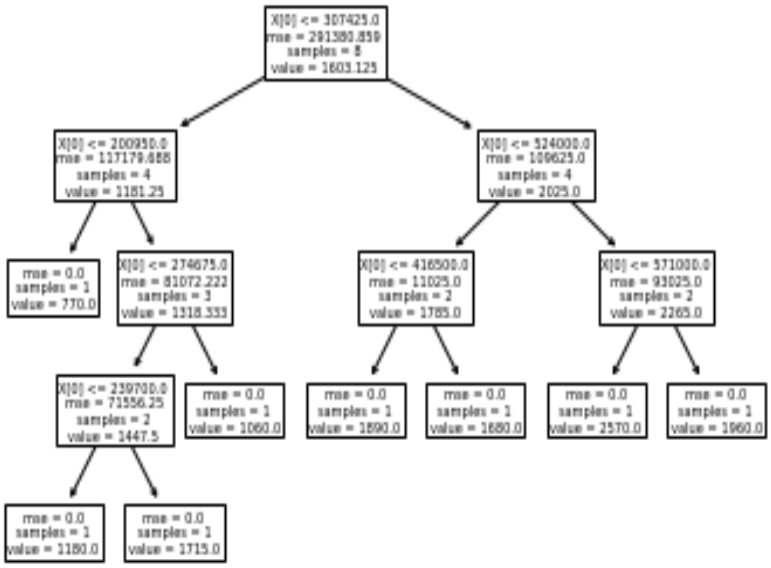

Fig.6.2.2 Tree implementation of price prediction using single attribute by decision tree

Similarly On applying the same example as we applied in above algorithms , we get the output 365000. which means if the customer have the requirements as we provide in our input then the cost of his desired house will be around Rs 365000.00 lac.
In linear regression , **RMSE value** is **205137.39**

6.3 Random-forest Regressor

In Random-forest regressor we are successful to predict the price with an accuracy of **96%.** Graphical representation of result on the basis of sqft_living:

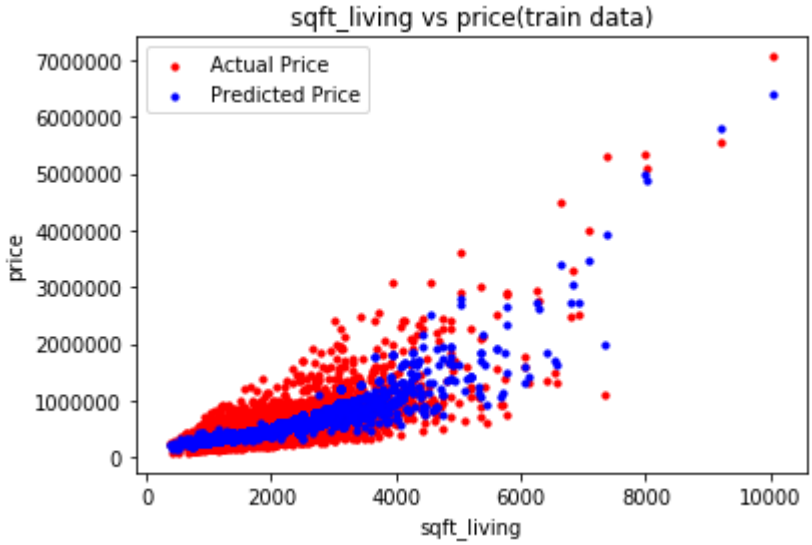

Fig.6.3.1 Price prediction on sqft_living by Random-forest regressor(Train data)

Fig.6.3.2 Price prediction on sqft_living by Random-forest regressor(Test data)

On applying the same example as we applied in above algorithms, we get the output 249280.76333. which means if the customer have the requirements as we provide in our input then the cost of his desired house will be around Rs 249280.76 lac.
In this ,**RMSE value** is **65650.14.**

7. CONCLUSION OF THE REPORT AND FUTURE SCOPE

In this report several tests have been performed using Various Algorithms (Linear regressor, Decision tree regressor & Random forest Regressor) to perform house price prediction. We achieved maximum accuracy with Random forest regressor algorithm i.e**., 96%.** Also the **RMSE**(root mean squared error) value in Random forest is minimum which proves the efficiency of algorithm.

This accuracy & efficiency of algorithm will help the real estate business to achieve correct price of the desired house in future. For Future work ,we will implement a working model with the help of website that predicts the house price on the basis of desired requirements given by investors.

8.REFERENCES

[1] Nissan Pow, Emil Janulewicz, Liu (Dave) Liu, 2016. Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal: An application of Random Forest, McGill University.

[2] Itedal Sabri Hashim Bahia, 2013. A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study ,International Journal of Intelligence Science, pp. 162-169 .

[3] Yu, Jiafu Wu. 2016. Real Estate Price Prediction with Regression and Classification, CS 229 Autumn 2016 Project Final Report, Stanford University.

[4] Da-Ying Li, Wei Xu, Hong Zhao, Rong-Qiu Chen, 2009.A SVR based forecasting approach for real estate price prediction,International Conference on Machine Learning and Cybernetics, Baoding.

[5] L.Li, K.-H. Chu, 2017. Prediction of real estate price variation based on economic parameters, 2017 International Conference on Applied System Innovation (ICASI).

[6] Fletcher M., Gallimore P. and J. Mangan, 2000, "The Modeling of Housing Sub-markets", Journal of Property Investment & Finance, 18(4).

[7] Owen C. and J. Howard, 1998, "Estimation Realisation Price (ERP) by Neural Networks: Forecasting Commercial Property Values", Journal of Property Valuation & Investment, 16(1): 71 – 86.

[8] Byeonghwa Parka Jae, Kwon Bae, 2014.Using machine learning algorithms for housing price prediction: The case of Fair-fax County, Virginia housing data: Expert Syst.Appl.42,2928-2934(2014).

[9] Tay D. P. H. and D. K. H. Ho, 1991, "Artificial Intelligence and The Mass Appraisal of Residential Apartments", Journal of Property Valuation & Investment, 10(2): 525 – 539.

[10] Worzala E., M. Lenk and A.Silva, 1995, "An Exploration of Neural Networks and Its Application to Real Estate Valuation", The Journal of Real Estate Research, 10(2): 185 – 201.