Real Estate Price Prediction Using Machine Learning

MSc Research Project
Data Analytics

Aswin Sivam Ravikumar
x16134621

School of Computing
National College of Ireland

Supervisor: Thibaut Lust

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Aswin Sivam Ravikumar
Student ID:	x16134621
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Thibaut Lust
Submission Due Date:	11/12/2017
Project Title:	Real Estate Price Prediction Using Machine Learning
Word Count:	6000

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Real Estate Price Prediction Using Machine Learning

Aswin Sivam Ravikumar
x16134621
MSc Research Project in Data Analytics

11th December 2017

Is it possible to predict the real estate house predictions effectively using Machine learning algorithms and advanced data mining tools.

Abstract

The below document presents the implementation of price prediction project for the real estate markets and housing. Many algorithms are used here to effectively increase the accuracy percentage, various researchers have done this project and implemented the algorithms like hedonic regression, artificial neural networks, AdaBoost, J48 tree which is considered as the best models in the price prediction. These are considered as the base models and by the help of advanced data mining tools algorithms like a random forest, gradient boosted trees, multi layer perceptron and ensemble learning models are used and prediction accuracy is attained in a higher rate. The results and evaluation of these models using the machine learning and advanced data mining tools like Weka, Rapid Miner will have the more influence in the price prediction.

Keywords: Random Forest,Multiple Regression,Support Vector Machine,Gradient boosted trees,Multi layer perceptron,bagging,price prediction,R,Weka,Rapid miner,Machine learning,Advanced data mining.

Contents

1 Introduction

We need a proper prediction on the real estate and the houses in housing market we can see a mechanism that runs throughout the properties buying and selling buying a house will be a life time goal for most of the individual but There are lot of people making huge mistakes in united states of America right now when buying the properties most of the people are buying properties unseen from the people they dont know by seeing the advertisements and all over the grooves coming around the America one of the common mistakes is buying the properties that are too expensive but its not worth it. In the housing market 2017, there is a survey that in the year 2016 the house sold in the America were about 5.42 million but the starter home inventory down up to 10.7% from 2015.

There was an economic collapse in the year 2007 and 2008 so there were several economic indicators that give the clue of impending disaster, this situation is currently happening and the economic indicators suggest that the housing prices are getting high people uses the real estate to known the current economic situations, the US government fails to produce the data about the house prices so it becomes difficult to buy the properties so the 87% of the people who needs to buy houses are using the Internets to search so there is evidence there is a correlation between housing sales and housing prices.

In general, real estate may have the valuation of land may be obliged to furnish. A quantitative measure of the profit is carried out by many different Players in the commercial center, for example, land agents, Appraisers, assessors, mortarboard lenders, brokers, Developers, gurus Also reserve managers, lenders,etc. Business worth will be evaluated through that requisition. From claiming valuation systems Also methods that reflect those nature Of property and the condition under which those provided for. The property might well on the way exchange in open market under many conditions and circumstances, people are the unaware amount the current situations and they start losing their money, the change in prices of properties would affect both the common people and the government, to avoid certain circumstances there is a need of price prediction.

Many methods have been used in the price prediction like a hedonic regression in this I am trying to predict the predict the real estate price for the future using the machine learning techniques with the help of the previous works. I have used the random forest, multiple regression and more algorithms with different tools to predict the house price So, it would be helpful for the people, so they will aware of both current and future situations, so it may avoid them in making mistakes. The remaining paper is organized as section 2 describes the previous works done by different researchers using different algorithms section 3 provides the methodology and the tools used and section 4 explains the way that the algorithms implemented, and comparisons and results are given in the last section.

2 Related Work

Before committing to the project several ground works should be done so there is a need for literature review I have analyzed many papers regarding the price prediction related the house markets and other different sectors. The papers I have taken will be in the different years up to the present year and I have used the recent and latest technologies, our main goal is to get more accuracy than the previous works the below passages will describe the past prediction works done by the various researchers and it will be helpful to implement the corresponding project.

2.1 A Review Of Price predictions

At present each framework may be moved towards innovation for the simplicity from claiming operations. The training framework will be moving towards e-taking. Individuals tend to move from the manual to robotized methodology. That primary goal of the this will be will anticipate that lodging cost with admiration to the plan of the clients. Those exhibit strategies may be An long procedure in which those customers necessities to contact the land operator. The land operators give acceptable A suggestive on the lodging costs prediction. This strategy includes high hazard a direct result the land operator might furnish the bad data of the clients. They employments those straight relapse calculations should figure the cost. This analyses likewise utilized to foresee the best area for the clients to purchasing the houses. The information here utilized is from those Mumbai lodging board since 2009. Eventually, Tom's perusing utilizing this straight relapse he predicted the rate for every square foot. This prediction indicates the square feet of the house will be raised Eventually Towards 2018.(Bhagat et al.; 2016)

The mankind's wealthiness is measured Eventually Purchasing a house includes a considerable measure from claiming consolidated choices. Concerning illustration, the same approach offering A house may be additionally troublesome. There needs aid where both the customers and sellers should get them an equal amount of profit. A different model will be carried Eventually that is the Cox regression model for those exact prediction. This model may be propelled starting with survival Investigation. That information utilized for this prediction is from the website named Trulia. He Additionally suggested that it is difficult to get the actual selling time with the website time this is because the observation time is far beyond the data. He also suggests that unsupervised learning methods have more popularity when compared to the other methods. (Li and Chu; 2017) also says that survival regression method helps to predict the values with the help of using the each and very attributes related to the house and its surroundings.

(Li and Chu; 2017)A late worth of effort carried out Toward to house value. The value of the house may be influence Toward Different budgetary factors. As we all know that China is one of the most populated countries around the world. Here the author tries to make a prediction to help the banks to provide the home loan for the customers. That prediction compares to Cathy house value list provided by the China. The information is gotten from Taipei lodging segment the over-proliferation after the data collection they use the machine learning algorithm neural network to predict the price the and accuracy of the prediction can find out using the RMSE (ROOT MEAN SQUARE ERROR) and the MAE(MEAN ABSOLUTE PERCENTAGE ERROR).(Li and Chu; 2017)(Willmott; 1981)

(Park and Bae; 2015)During the year 2005 there seems to be a high rise in interests on the American housing markets so the America was forced into bankruptcy so it reflects US housing markets they were declined up to 30 to 60 % in the major cities it was continued for many years, after the November 2012 it started to recover because the investment becomes low so there was a demand so the author tries to research and developed a prediction model to get whether the closing price is higher or lower by using the machine learning to obtain the knowledge and to predict the future. Here he uses the KDD model knowledge discovery databases the data here used seems to be merged from the different data sets and uses the WEKA software to find them a multiple algorithms like decision tree is used to finding the relationship in the database, here park and bay uses the RIPPER, C4.5 (J48), naive Bayesian and Ada-Boost every algorithm is used under different conditions RIPPER is used for selecting the majority class and minority class, naive Bayesian is used to divide the data set into different classes by calculating the probability distribution and AdaBoost is used to improve the classification and here performs the two methods one is three way split with 10 folds and 10 folds cross validation by his results achievements RIPPER have the more prediction compared to the other.

(Piazzesi and Schneider; 2009)Those foreseeing those value of the product alternately an arrangement may be altogether intricate. The cost prediction is basically utilized within impart business sector. Yet the prediction from claiming offer worth may be precise perplexing due to it dynamic clinched alongside the way. Need to be carried out a neural system model to foreseeing the stock value. This gives an association between those stocks Also benefit. In this model, the creator utilized the stock information need. The following venture will be to ascertain those relapse components based upon the shutting esteem of the stock. Straight relapse will be performed on the first information situated. Right away the duplicate of the first information situated is made What's more Fourier analysis may be connected. Following that Fourier analysis, that standardization of information will be finished. This makes an MLP with a portion neuron. Right away the neural system calculation may be actualized. This calculation gives preferred correctness done prediction Also offers great commotion tolerance. The principal hindrance is that the stake business sector information continues overhauling and the prediction turns into was troublesome.

The author(Gu et al.; 2011) says that housing price involves the various economic interest it also includes both the government and the peoples so there is in need of accurate forecasting so the three researchers from key laboratory developed a new model using the genetic algorithm and the support vector machine. They have clearly mentioned the regression theorem of the support vector machine and introduced a new function called kernel with the help of Karush-Kuhn-Tuckers(KTT) conditions. here they have combined the genetic algorithm with the SVM and named it as G-SVM where the kernel functions will be in chromosomes and each will divide into three segments the author is aware of the fitness model so they have calculated the fitness value of for each chromosome so there will less percentage of over fitting model and three operations selection, crossover and mutation operation are performed and the results are obtained . there is a comparison between the grey model and GSVM and GSVM executes the results faster and more precise as suggested by the founders.

The authors(Limsombunchai; 2004) try to provide a more accurate prediction on the house prices to improve efficiency to the real estate present in New Zealand he suggests that most peoples in New Zealand have their own houses the sample data is obtained from one of the trusted real estate agency so we can believe that there will not be an error in the data here he compared the hedonic price and the artificial neural network theory when conducting the hedonic price model there is hypothesis based on the previous works it seems to have a positive relationship. In the neural;l network the author uses the trained data in order to avoid the prediction errors the work strategy of neural networks is stated clearly and at last when comparing the results the author says that artificial neural network has more performance when compared to the other one.

The author determines the housing prices in the turkey as by the method (Limsombunchai; 2004) the data is taken from the household survey during the year 2004 they suggest that hedonic multiple regression models are mostly used for the price prediction because they tend to fit the data into the model by observing the results there is no multicollinearity between the variables but there is heteroscedasticity due to the white state statistics suggesting that there will be potential problem in the model also some variables are dominating the significant variables in the house price predction while coming to the artificial neural networks they suggest that networks have the capacity of adapting and it is also one the flexible models in this predection feed forward network is used for prediction by comparing their results ANN (ARTIFICIAL NEURAL NETWORK) tends to have more performance when it compared with the hedonic multiple regression as suggested by the(Nghiep and Al; 2001)

The authors(Selim; 2009) have compared the multiple regression analysis over the artificial neural networks by using the 60% data for the house pricing prediction several comparisons have been made in their predictive performance they have compared with the different training size and selecting the data in their size ie) the sample data size various for the performance detection. For calculating the error two different equations are used mean absolute percentage error and the absolute percentage error, here the absolute percentage divides the properties into three different stages based on the FE(Forecasting error) percentages Totally six different comparisons are made for more efficiency, here it's clear that if there is enough or sufficient data size artificial neural network can perform better or else the results will be different as said by(Willmott; 1981)

The two authors (Wu and Brynjolfsson; 2009) from MIT have conducted about the prediction that how the Google searches the housing price and sales across the world suggesting that in the present world every prediction percentage point is correlated with the next year house sales. The author reveals about the correlation between them housing price and their related searches and the positive relationship between them. The data is taken from the Google search which means the search queries by using the Google trends and with the help of a national association of real-tors the data is collected for all the states present in the united states of America and found the highest number of houses sold during the year 2005 and the recession starts over 2009 by using the auto regressive (AR) model, by using it the relationship between the search queries and housing market indicators they have estimated the baseline for housing price prediction and they are well shown in the figures and suggesting that if there demand to house and there will be demand in house hold appliances.

The author gives the brief detail about how the random forest algorithm is used for the regression and classification, boosting and bagging are said to be the methods which produce a many classifiers the difference between the boosting and bagging Is as said by (Liaw et al.; 2002) is the successive tress, the point weights are calculated and majority will take for the prediction.during the year 2001 (Nghiep and Al; 2001) he proposed the random forset which is related to bagging and it gives more randomness the complete process of random forset classification and regression are stated here for the regression they have used the Boston housing data and find out the forecasting error ie) RMSE root mean squared value as

$$MSEOOB = n - 1 \; n \sum 1 \; \{yi - y\hat{}\; OOB \; i\}| \hat{}2$$

Figure 1: MSE

here the author speaks about the variable importance and the ocean proximity, variable importance helps to give whether the prediction error increases or not and also which variable will have more influence on the prediction and the proximity error helps to find shape and structure of the data.

(Breiman; 1996)Bagging predictors a method which helps to give various versions of a predictor with the support of aggregated predictor the bagging algorithm is used in both moderate and large data sets on par as for both regression and the classification as same the Boston housing data is used for the here the data is divided in the ratio of 10:90 10 % of the testing and the remaining 90% for the training and by using the 10 fold cross-validation regression tree is built and the process takes places for this data testing and the training process is continuously takes place over 100 times for each time a new cases will be generated and the RMSE estimated standard errors seems to be decreased which the prediction accuracy is high when it compared with the other.

For classification problems, there is a way to find out the accuracy percentage with the help of the confusion matrices we can find out the accuracy percentage but the regression there is only one possibility to calculate the RMSE root mean squared error here the author says about the error indices an average error of a model calculated using the mean squared error (MSE) or the root mean squared error (RMSE) . there is a problem in using the correlation coefficient as the significance test, not an appropriate one so that we prefer the RMSE.

2.2 Summary of Literature Review

For every project the literature review will give clear idea and it will serve as the base line here most of the authors have concluded that artificial neural networks have the more influence in predicting but in the real world the other algorithms should be also taken into consideration. by conducting this study it helps to know about both the pros and corns and it had helped me to successfully implement the project.

3 Methodology

The below passages describe about the methodology used in the real estate house price predictions and the architecture flow diagram is given.

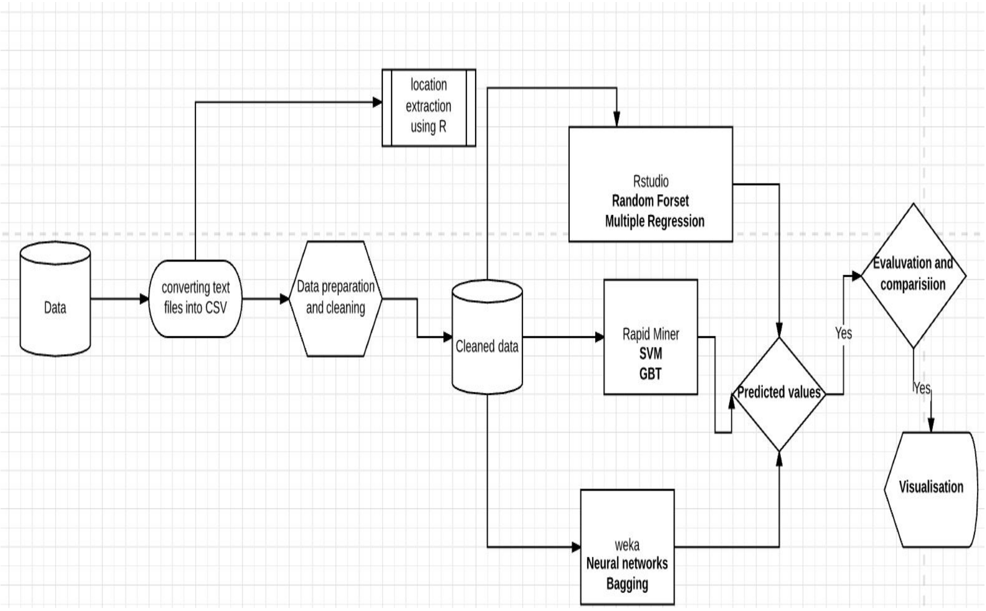


Figure 2: Architecture of price prediction

3.1 Description of Data-sets

The real estate housing data is used in this and it is taken from the UCI machine learning repository and the ageron the data is spread across 20000 rows and has the ten attributes the description of the data set is given below

variable table		
S.no	**Variables**	**Integer type**
1	Latitude	Real
2	Longitude	Real
3	Housing median age	Integer
4	Total Rooms	Integer
5	Total Bedrooms	Integer
6	Population	Integer
7	Households	Integer
8	Median Income	Real
9	Median House house(Price)	Integer
10	Ocean Proximity	Poly-nominal
11	Special Attribute Y	value to be predicted

Here there are totally 10 predictor variables and the Y variable will be median house price which is going to be predicted.

3.2 Data Cleaning And integration

The data obtained from the repository is in the form text file I have connected the text with the excel and the data is being extracted from the text file and moved into the excel file and it has been saved as the comma-separated file. Data cleaning is an iterative process, the first iterate is on detecting and correcting bad records the data taken from the repository have many inconsistencies and null values before loading into the machine learning models the data should be corrected in order to get the high accuracy of prediction as I am using the different tools for prediction, the cleaning process differs from one other but the ultimate goal is to gain more accuracy. The real estate data have some missing information they dont have the states name only latitude and longitude were given by using the R program I have identified the states they all seem to be the states present in the united states of America and the null values are removed to reduce the inconsistency.

3.3 Detection Of Outliers

An outlier is an extremely high or extremely low-value value in the data it can be identified if whether the value is greater than interquartile range Q3 + 1.5 or Q1 - 1.5 detecting the interquartile range is arrange the data in an order from the lower value to the higher value, now the mean is taken for the first set of values and second set values now by subtracting both mean we can get the interquartile range the formula is Q3 + (1.5)(quartile range) and for Q1-(1.5)(quartile range) and I have calculated using the R program.

3.4 Tools

Here there are some essential tools used for the prediction project.

Tools and Algorithms		
S.no	**Name of the tool**	**Algorithms used**
1	R studio	Random forset,Multiple Regression
2	Rapid Miner	Support Vector Machine,Gradient Boosted Trees
3	Weka	Neural networks,Bagging

3.4.1 R Studio

Its one of the open source and free development tool used for the statistical, machine learning and graphics tool.

3.4.2 Rapid Miner

Rapid miner is a data science platform and its one of the open source innovation integrated with many analytical methods and the predefined machine learning algorithms the reason for using them is to find whether the tools have any significant influence in giving more accuracy when compared to the written programs.

3.4.3 Weka

Weka is graphical interface learning which computes the machine learning and data mining techniques the working process differs from every different platform here the algorithm can be directly applied to the model with the several conditions.

3.4.4 Regression

It is a data mining task of predicting the value of target(numerical variable) by building a model based on the one or more predictors the predictors can either be numerical or the categorical variables.

3.5 Machine Learning Algorithms

3.5.1 Random Forset

(Patel et al.; 2015)Random forest algorithm can be used to predict both the classification and the regression it is also be called as the regression forests. The main process is it develops lots of decision tree based on the random selection of data and the random selection of variables and it provides the class of dependent variable based on many trees. The main advantage of using this algorithm to my dataset is it handle the missing values and it can maintain the accuracy of the missing data and the chance of overfitting the model is low and we can except high dimensionality when we apply to the large level dataset. In regression trees, the outcome will be continuous.

3.5.2 Multiple Regression

Its a new version of the linear regression which is considered to be more powerful which works with the multiple variables or the multiple features it helps to predict the unknown value of the attribute from the known value of the two or more attributes which will be also known as the predictors (Chang and Liu; 2008)

3.5.3 Support Vector Machine

Support vector machine regression is derived from the classification algorithm known as the support vector machine SVM produces a hyperplane that separates the points with the different labels here I am using the similar method instead of separating the data it produces a hyperplane that is close to the most of the points. So the price can be predicted by (Trafalis and Ince; 2000)

3.5.4 Gradient Boosting

As said by(Ganjisaffar et al.; 2011)Gradient boosting can be used for both the regression and classification gradient boosting is a technique for producing regression models consisting of collections of regressors it is an instantiation of this idea for regression the main theme is to repeatedly follow the procedure here we are learning the simple regression predictor of the data then we are computing the error residual. The amount of the error per data point and we a learn a new model to predict the error residual. The main concept is we are making a set of predictions then finding the errors and we are reducing it.

3.5.5 Neural Networks

MLP is the multilayer perceptron it is a part of artificial neural networks it has the same structure of a single layer percepron with one or more hidden layers, in this the hidden layer will be directly connected to the input layer here the input values are presented in the perceptrons, the perceptrons will classify any linear separable set of inputs if the input values presented to the perceptron, and if the predicted output is as same as the desired output, then the performance is considered satisfactory and we know that no changes to the weights are made, if it does not match then the weights need to be changed to reduce the error.(Koskela et al.; 1996)

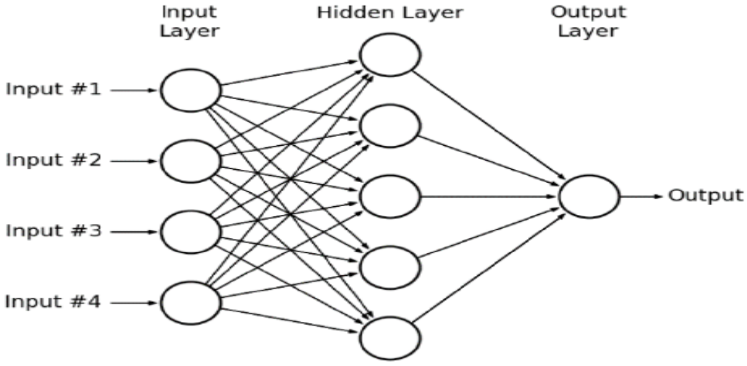


Figure 3: Multi layer perceptron Model

3.5.6 Ensemble learning Bagging

(Mirmirani and Cheng Li; 2004)It is part of machine learning known as the ensemble learning mete algorithm which is designed to improve the stability and to reduce the variance and accuracy. Its nothing but the application of group learning so essentially here multiple models are built they come together and bring a forth model that is more accurate. In bagging various models are built in parallel on various samples and then the various models vote to give the final model and hence prediction.

3.5.7 Accuracy Calculation And visualisation

The indicators I am used to evaluating the performance accuracy is the mean absolute error which is the difference between the predicted value and the actual value. After prediction the results data will be loaded into the tableau so it can be clearly visualized and it can also be used for the future works.

4 Implementation

The main aim of this project to be implemented is to find out the accurate prediction of the real-estate properties present in the united states of America for the next upcoming years, the below segment blankets will help you to know the implementation process in depth. Here step by step process involved is represented below.

1. Scientific Environment.
2. Source of a Data.
3. Excel 2016: the first process to store the data.
4. Loading data into R, Rapid Miner, Weka.
5. Normalizing the data.
6. Detecting Outliers.
7. Analysis and visualization using the R, Rapid Miner, Weka.
8. Machine learning models are build using the Cat Tools, and the various algorithm used for predictions as listed in the methodologies.
9. Splitting the data sets as test and train for the Cross-validation process.
10. Fitting the data into different machine learning and data mining models for the Predictions.
11. Finding the Root mean square value and R square to finding the Accuracy percentage.
12. Visualization.

4.0.1 Need of a Technical Environment

1. Microsoft Excel
2. R studio for creating the Scripts
3. Essential Libraries to be installed in R
4. Rapid Miner
5. Weka
6. Appropriate Functions to be selected in both the Rapid Miner and Weka
7. Tableau For visualization

4.0.2 Data Source

As I mentioned before the main data sets were collected from the UCI machine learning repository which is open data resource for the data mining and for the predictive analytics purposes. The acquired data source was a text file for finding the errors in the data source, the text file is connected to the EXCEL and they are separated using the commas and saved as a CSV file.

4.0.3 Data cleaning

The data in the CSV files need to be checked whether it has any missing values, as the data source have missing values every attribute have checked using the filters and null values are removed which helps to increase the accuracy level.

4.0.4 Tool 1 R studio

The Real estate data is processed with the help of R program in R studio and the numerical data is normalized using the script so the data will equally distributed and variables cannot dominate each other. It helps to bring the values of the attribute on a common scale.

4.0.5 Implementation of Random Forset and Multiple Regression

Packages Used: Random Forest, nlme, Cat Tools
The data is been split into test and train 75 percent data is used for the test data and remaining for the train data since I am using the Random forest the number of trees used is 200.Consider the standard recursive partitioning algorithm will start it searches all the data and in-depth search is made for all the variables and the best-split point is taken and the process gets repeated for the right and left leaves. Here in the random forest, the variables are selected randomly from MTRY variables with the help of predictor attributes for each split the different variables are selected. we can find the variable importance, so here the median house value has the more influence in predicting. The prices have been predicted and its shown in the graph. Here the average output of several trees has been taken in order to provide more accuracy.

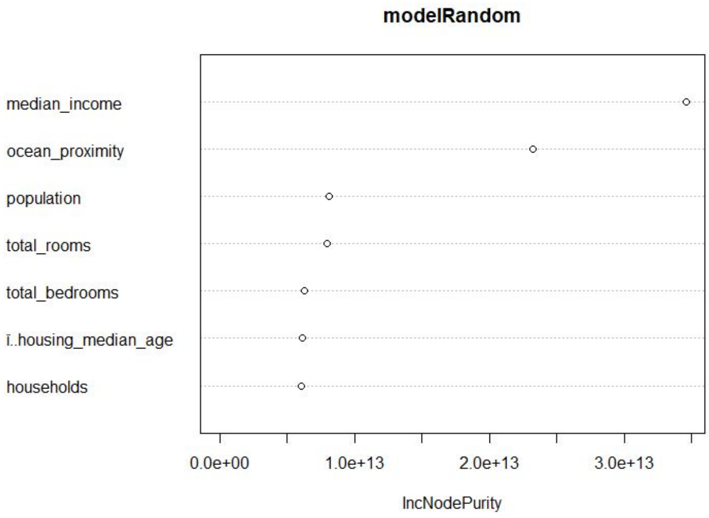


Figure 4: feature extraction(variable importance)

The same method has been followed for multiple regression with the different assumption here the dependent variable is a median house value attribute which is going to be predicted and the other attributes are independent variables and the attributes are given into the model and with the help of coefficients an equation has been made for regression and the predicted prices are produced.

4.0.6 Tool 2 Rapid Miner

The data source is added to the repository of a rapid miner, the tool has the inbuilt functions to remove the null values and it is been stored in the data repository. Functions Used: Data source, Select attributes, Set Role, Normalize, Cross Validation Training, Testing.

4.0.7 Implementation Of SVM and Gradient Boosted

In rapid miner, the algorithms are the predetermined one the data loaded is connected to the function named select attributes here the not null values have been taken in order to avoid the null values inside the prediction process, every attribute will have only one role in order to predict the attribute with the help of set role function the role of the median house value is been changed into the label and connected to the normalize function since the price ranges are too high this variable has the more chance in dominating the other variables so its been normalized to a particular scale value and sent into cross validation. The cross-validation has the two parts one is training and the next is testing the data set. The basic idea of cross-validation is to partition the data set into k bins of equal size so according to the bins the data will be separated here I am using 10 fold cross validation 9 parts of the data will be used as testing data and the remaining one part will be used to the train the data as the algorithms we need, I have used the gradient boosted tree and the support vector machine. In the testing part, both the data can be predicted and the performance also can be classified. The performance support vector count helps to produce the predicted variables and the performance regression will help to give the RMSE root mean square error values, here the prices have been successfully predicted as and the error percentages are discussed in the evaluation part.

4.0.8 WEKA implementation of Neural networks and Bagging

The data loaded into weka is changed to arff format and the is been normalized and saved as a new file, and the normalized file will be loaded here also the 10 fold cross validation used but for the neural network multilayer perceptron and bagging here 70 percent of data is used for testing and 30 percent for training in general multilayer perceptron will have three layers input, hidden and output, as I discussed in methodologies, the weighted average of the input layer, is sent to the hidden layer and the output layer. here the all the independent variables like households, population, income, bedrooms are fed into the input layer and the middle layer are neurons consisting the average weight and the third layer is the output layer which gives the predicted price. The process is repeated for bagging in the input variables that the independent variables help to generate the multiple input data the test data and train data percentage remains same as the multilayer perceptron by generating the multiple inputs there is a regularization and there is a chance to have the more accuracy. The predicted values and The estimation of errors is discussed in the evaluation part. .

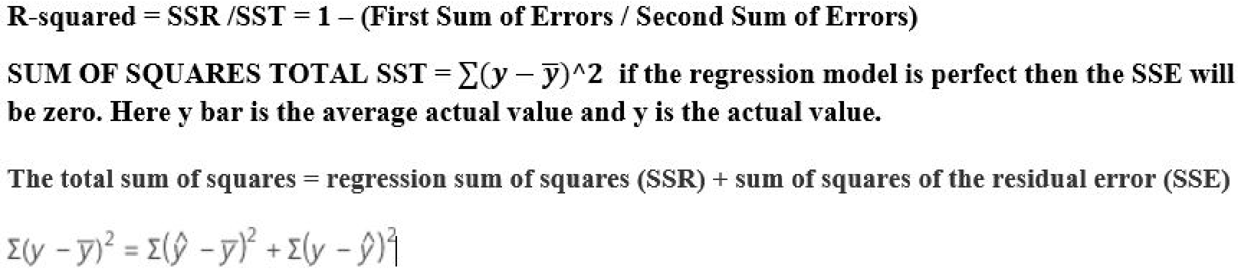
5 Evaluation

The idea of a regression is to predict a real value which means number in regression model we can compute the several values the most common terms are explained below

5.1 Coefficient of determination

The coefficient of determination R square summarizes the explanatory power of the regression model and is computed from the sum of squares terms.

The R square describes the proportion of variance of the dependent variable explained by the regression model and the equation is given below



R-squared = SSR /SST = 1 – (First Sum of Errors / Second Sum of Errors)

SUM OF SQUARES TOTAL SST = $\sum(y - \overline{y})$^2 if the regression model is perfect then the SSE will be zero. Here y bar is the average actual value and y is the actual value.

The total sum of squares = regression sum of squares (SSR) + sum of squares of the residual error (SSE)

$$\Sigma(y - \overline{y})^2 = \Sigma(\hat{y} - \overline{y})^2 + \Sigma(y - \hat{y})^2$$

Figure 5: Coefficient of determination

The figure shows the correlation between my attributes and how they are related with each other

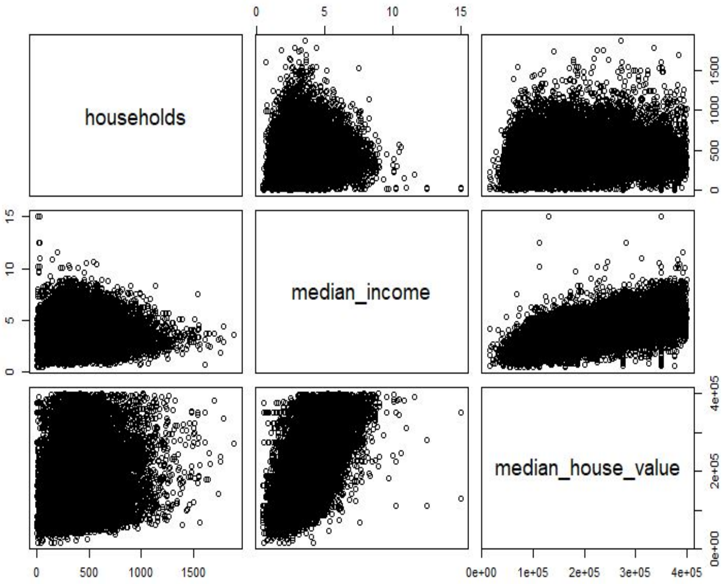


Figure 6: Correlation between attributes

5.2 Root Mean Square Error

RMSE is a popular formula to measure the error rate of a regression model, however, it can only be compared between models whose errors are measured in the same units it can be measured using the given formula

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)^2}{n}}$$

Figure 7: rmse formula

Where n is the number of instances in the data, P is the predicted value for the I instance and O is the actual value the key concept is the predicted value is subtracted by the actual value square that and get the sum of all instances and divided it by number of instances, the RMSE will be achieved.

As discussed, the essential variables are used to calculate the error value and help to determine the how well can the algorithm predict the future prices, the below table describes the Root mean square error for the various algorithms and displayed below.

5.3 Evaluation of Result

RMSE			
S.no	**Algorithms**	**RMSE Error**	**Accuracy**
1	Random forset	0.012	90%
2	Neural Network	0.590	60%
3	Gradient Boosted	0.573	65%
4	Bagging	0.563	70%
5	Support Vector machine	0.636	58%
6	Multiple Regression	0.70	55%

5.4 Over all Case Study

The first case study describes the results of random forset and multiple regression, below the following graph shown the actual vs predicted value

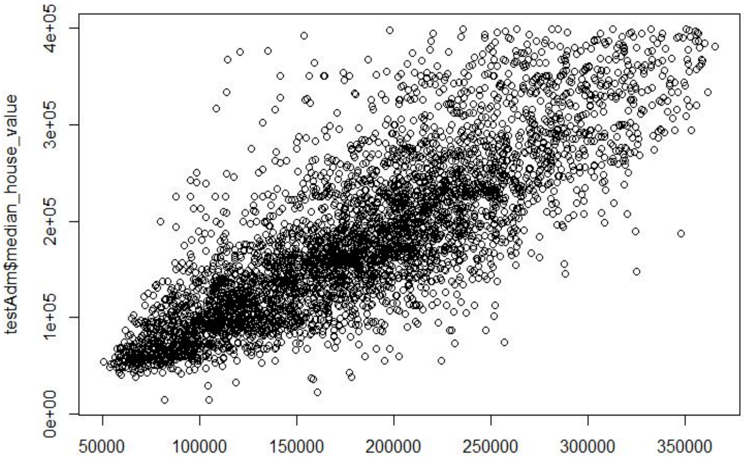


Figure 8: Random forset Predicted vs Actual

Here the scatter plot shows the actual and predicted values are spread across linearly so that the price predicted is accurately about 90% with the RMSE value of 0.012.

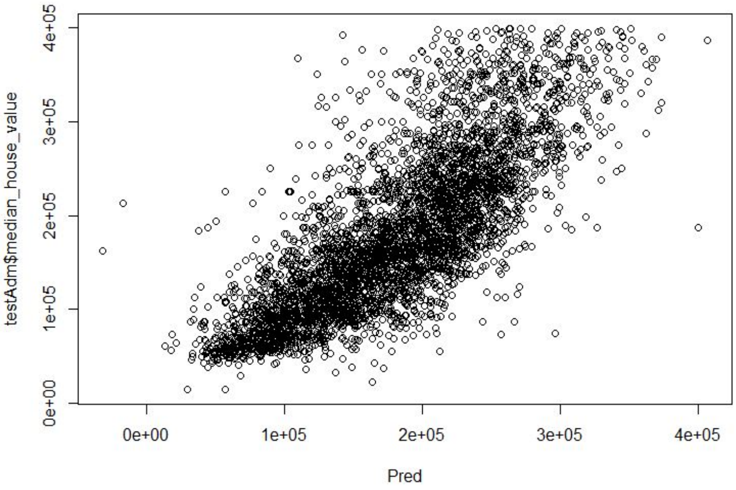
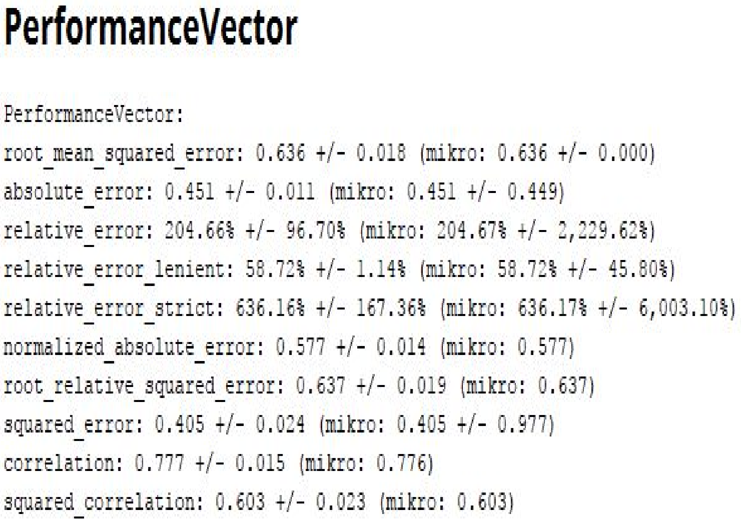


Figure 9: Multiple Regression Predicted vs Actual

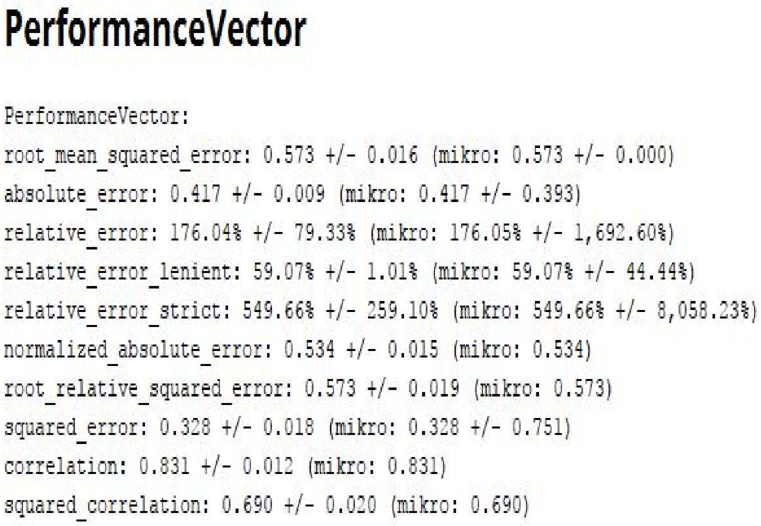
The above scatter plot for the multiple regression shows that values are not spread-ed linearly so the the actual value and the predicted price value are not similar but the predicted percentage is about 55% with the RMSE value of 0.701.



PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.636 +/- 0.018 (mikro: 0.636 +/- 0.000)
absolute_error: 0.451 +/- 0.011 (mikro: 0.451 +/- 0.449)
relative_error: 204.66% +/- 96.70% (mikro: 204.67% +/- 2,229.62%)
relative_error_lenient: 58.72% +/- 1.14% (mikro: 58.72% +/- 45.80%)
relative_error_strict: 636.16% +/- 167.36% (mikro: 636.17% +/- 6,003.10%)
normalized_absolute_error: 0.577 +/- 0.014 (mikro: 0.577)
root_relative_squared_error: 0.637 +/- 0.019 (mikro: 0.637)
squared_error: 0.405 +/- 0.024 (mikro: 0.405 +/- 0.977)
correlation: 0.777 +/- 0.015 (mikro: 0.776)
squared_correlation: 0.603 +/- 0.023 (mikro: 0.603)
```

Figure 10: SVM RMSE



PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.573 +/- 0.016 (mikro: 0.573 +/- 0.000)
absolute_error: 0.417 +/- 0.009 (mikro: 0.417 +/- 0.393)
relative_error: 176.04% +/- 79.33% (mikro: 176.05% +/- 1,692.60%)
relative_error_lenient: 59.07% +/- 1.01% (mikro: 59.07% +/- 44.44%)
relative_error_strict: 549.66% +/- 259.10% (mikro: 549.66% +/- 8,058.23%)
normalized_absolute_error: 0.534 +/- 0.015 (mikro: 0.534)
root_relative_squared_error: 0.573 +/- 0.019 (mikro: 0.573)
squared_error: 0.328 +/- 0.018 (mikro: 0.328 +/- 0.751)
correlation: 0.831 +/- 0.012 (mikro: 0.831)
squared_correlation: 0.690 +/- 0.020 (mikro: 0.690)
```

Figure 11: GBT RMSE

As mentioned in the evaluation the support vector machine has the RMSE about 0.636 with the 58% of accuracy and the gradient boosted tress have 0.573 RMSE and 70% of accuracy and the correlation between the atrributes and the other respected measures can be seen over the figure.

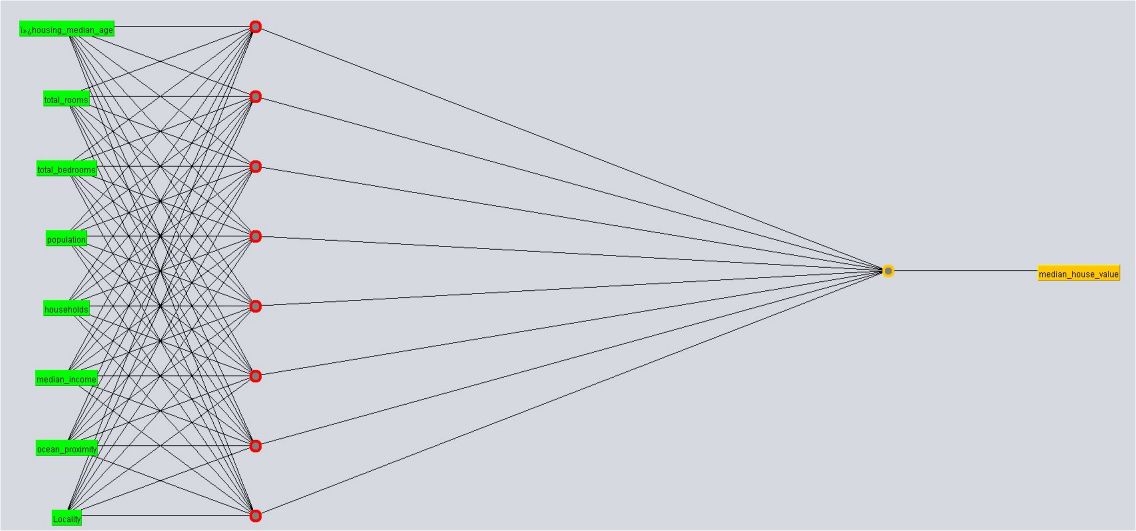


Figure 12: Multi layer perceptron

The figure 12 describes the input layers connected with the hidden layers, the RMSE for neural networks is about 0.590 and the accuracy percentage will be about 65% on par the bagging has the RMSE 0.53 and the accuracy will be 60%.

5.5 Tableau Visualization

Here the test data and the predicted price are loaded into the tableau and its clear that there is a linear trend between the actual and predicted values so the random forset model performs well and the accuracy percentage is more when we compared with the others.

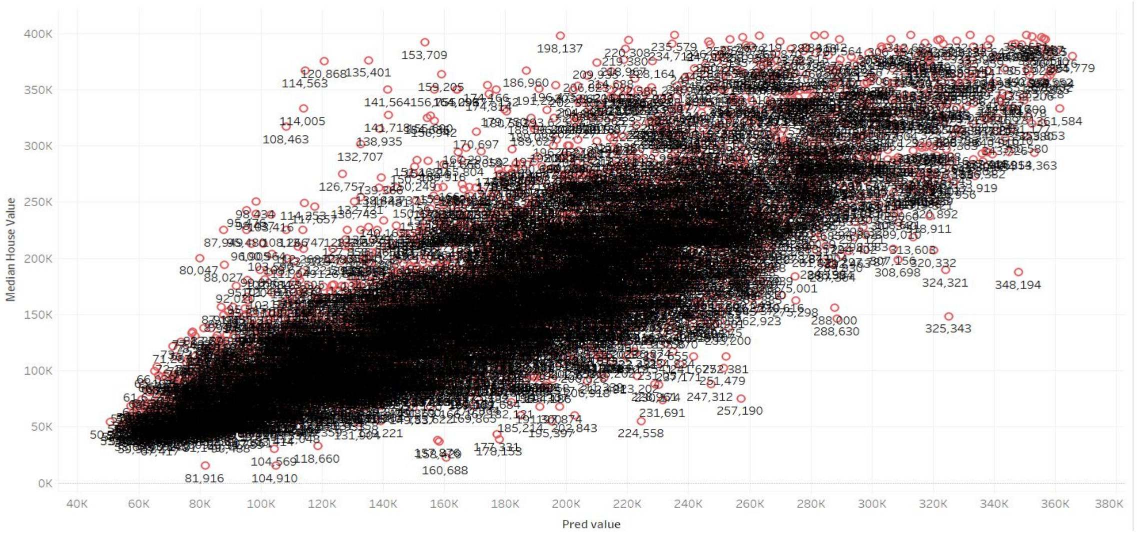


Figure 13: Test data vs predicted price

5.6 Discussion

By conducting this experiment with various machine learning algorithms its been clear that random for-set and gradient boosted tress are performing better with more accuracy percentage and with less error values. when this experiment is compared with the and to the result achieved these algorithms predicts well.this project has been done with the big data and its related technologies.

6 Conclusion and Future Work

The main goal of this project is to determine the prediction for prices which we have successfully done using different machine learning algorithms like a Random forest, multiple regression, Support vector machine, gradient boosted trees, neural networks, and bagging, so it's clear that the random forest have more accuracy in prediction when compared to the others and also my research provides to find the attributes contribution in prediction. So I would believe this research will be helpful for both the peoples and governments and the future works are stated below

Every system and new software technology can help in the future to predict the prices. price prediction this can be improved by adding many attributes like surroundings, marketplaces and many other related variables to the houses. The predicted data can be stored in the databases and an app can be created for the people so they would have a brief idea and they would invest the money in a safer way. If there is a possibility of real-time data the data can be connected to the H2O and the machine learning algorithms can be directly connected with the interlink and the application environment can be created.

7 Acknowledgements

I would thank my supervisor Mr Thibaut lust for providing his time in guiding me and the knowledge he shared with me, his guidance is the main reason to complete and successfully implement this project.

References

Bhagat, N., Mohokar, A. and Mane, S. (2016). House price forecasting using data mining, *International Journal of Computer Applications* **152**(2): 23–26.
URL: *http://www.ijcaonline.org/archives/volume152/number2/26292-2016911775*

Breiman, L. (1996). Bagging predictors, *Machine learning* **24**(2): 123–140.

Chang, P.-C. and Liu, C.-H. (2008). A tsk type fuzzy rule based system for stock price prediction, *Expert Systems with applications* **34**(1): 135–144.

Ganjisaffar, Y., Caruana, R. and Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models, *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, ACM, pp. 85–94.

Gu, J., Zhu, M. and Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine, *Expert Systems with Applications* **38**(4): 3383–3386.

Koskela, T., Lehtokangas, M., Saarinen, J. and Kaski, K. (1996). Time series prediction with multilayer perceptron, fir and elman neural networks, *Proceedings of the World Congress on Neural Networks*, INNS Press San Diego, USA, pp. 491–496.

Li, L. and Chu, K.-H. (2017). Prediction of real estate price variation based on economic parameters, *Applied System Innovation (ICASI), 2017 International Conference on*, IEEE, pp. 87–90.

Liaw, A., Wiener, M. et al. (2002). Classification and regression by randomforest, *R news* **2**(3): 18–22.

Limsombunchai, V. (2004). House price prediction: hedonic price model vs. artificial neural network, *New Zealand Agricultural and Resource Economics Society Conference*, pp. 25–26.

Mirmirani, S. and Cheng Li, H. (2004). A comparison of var and neural networks with genetic algorithm in forecasting price of oil, *Applications of Artificial Intelligence in Finance and Economics*, Emerald Group Publishing Limited, pp. 203–223.

Nghiep, N. and Al, C. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks, *Journal of real estate research* **22**(3): 313–336.

Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data, *Expert Systems with Applications* **42**(6): 2928–2934.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques, *Expert Systems with Applications* **42**(1): 259–268.

Piazzesi, M. and Schneider, M. (2009). Momentum traders in the housing market: survey evidence and a search model, *Technical report*, National Bureau of Economic Research.

Selim, H. (2009). Determinants of house prices in turkey: Hedonic regression versus artificial neural network, *Expert Systems with Applications* **36**(2): 2843–2852.

Trafalis, T. B. and Ince, H. (2000). Support vector machine for regression and applications to financial forecasting, *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, Vol. 6, IEEE, pp. 348–353.

Willmott, C. J. (1981). On the validation of models, *Physical geography* **2**(2): 184–194.

Wu, L. and Brynjolfsson, E. (2009). The future of prediction: How google searches foreshadow housing prices and sales.