JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

Website: http://www.jmrr.org Vol.3, Iss.1 Pages:16-22

ACCURACY ANALYSIS OF HOUSE PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

S Ponni Jenet[1], Dr P Revathi[2]

[1]II M.Sc Computer Science, [2]Assistant Professor, PG Department of Computer Science, Holy Cross College, Trichy, Tamilnadu, India

Article Received: January 2022 Published: April 2022

Abstract

House price predictions are expected to help people who are planning to buy a home by allowing them to know the price range in the future so that they may properly arrange their finances. House price predictions are also useful for property investors who want to know the trajectory of housing prices in a specific area. When it comes to buying a new home, people are cautious about their budgets and market strategy. The goal of this study is to forecast consistent house prices for non-home owners based on their financial resources and objectives. House prices are estimated by analyzing the following items, tariff limits, as well as other Predicts improvements. Predictions using machine learning algorithms such as Linear Regression and Random Forest are compared in the study. These algorithms are used to create house price prediction on Data set in order to determine the best among them. The conceptual goal of this paper is to assist and buy houses in the ways that they need to be protected against fraudulent brokers and excessive commissions.

Keywords: House Price Prediction, non-householders, Linear Regression, Random Forest

INTRODUCTION

Every year, house prices increase, necessitating the development of a method to predict future house prices. House price prediction can assist a developer in determining the selling price of a home and can also assist a client in determining the best time to buy a home. The main purpose of the paper is to estimate the price of a house based on its features. Finding the right algorithms and determining which one is best for predicting prices with a low error rate.

We attempt to demonstrate the many machine learning techniques that are applicable for our situation in this paper.

Linear regression is a supervised learning Machine Learning algorithm: The function of linear regression is to predict a dependent variable (target) based on a single independent variable(s)[1]. As a result, this regression technique determines if a given variable and other independent variables have a linear relationship.

A random forest is a machine learning algorithm used for supervised learning: This is used to solve problems involving regression and classification. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers.

PROPOSED SYSTEM

Dataset

Kc_house_data used the regression model to evaluate household data. Houses based on the dataset's features. Basic implementation steps are included. Data gathering and analysis are tools for better understanding datasets and identifying features in it. The data pre-processing stage cleans the database in preparation for sample development [2]. The train and test splits present two randomly formed subsets of our data; these test/train data are utilised to fit the learning algorithm and teach it how to predict. The model is then created and compared using the random forest and linear regression techniques that have been proposed.

The kc_house_data dataset was used to create the model. The dataset was made up of past data from home sales from May 2014 to May 2015. In kc house data, there are ten features for residences in various suburbs. Id, date, price, bed rooms, bath rooms, sqft living, sqft lot, floors, waterfront, views are all included in the dataset. The data set must be divided into the training dataset and the testing dataset because the model utilizes a supervised learning method. Eighty percent of the training data set was used to train the model, while the remaining twenty percent was used for testing.

The proposed model was developed by comparing the random forest and linear regression techniques. The RandomForestRegressor from the Python Sk-learn machine learning library was used to create the random forest. Linear regression was implemented using the Python Sk-learn machine learning library's LinearRegression. Both is a popular machine learning technique for supervised classification and regression.

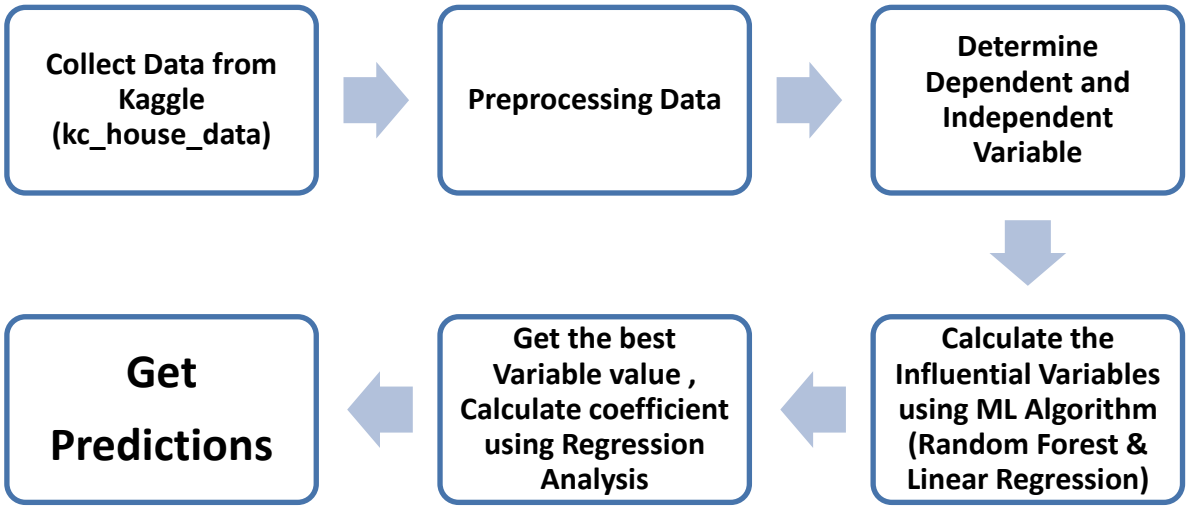

Fig 1 Data flow of Test, Train and Influential variables

IMPLEMENTATION

Hardware Requirements: For implementing this approach, Microsoft Windows 10 Operating System is used that includes Processor Intel Core i3 7020U at 2.30GHz and 4GB RAM.

Step 1: Import numpy package. If we predict the value using Linear Regression algorithm import the class LinearRegression from sklearn.linear_model.

```
In [2]: import numpy as np
        from sklearn.linear_model import LinearRegression
```

If we predict the value using Random Forest algorithm import the class RandomForestRegressor from sklearn.linear_model.

```
In [11]: import numpy as np
         from sklearn.ensemble import RandomForestRegressor
```

Step 2 : Load the kc_house_data Dataset

```
In [3]: import pandas as pd
        df=pd.read_csv("kc_house_data.csv")
```

Step 3 : Divide the data into independent and dependent variables

```
In [15]: y=train_data['price']

In [16]: x=train_data.drop(['id','price'],axis=1)
```

Step 4 : Split the data into train and test sets

```
In [17]: from sklearn.model_selection import train_test_split
         x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.20,random_state=42)
```

Step 5 : Train the Linear Regression algorithm using this step

```
In [20]: from sklearn.linear_model import LinearRegression
         linear_reg=LinearRegression()
         linear_reg.fit(x_train,y_train)

Out[20]: LinearRegression()
```

If we want to train the Random Forest algorithm using this step

```
In [22]: clf = RandomForestRegressor(n_estimators=300)

In [23]: clf.fit(x_train, y_train)

Out[23]: RandomForestRegressor(n_estimators=300)
```

Step 6 : Finally test the accuracy value for Linear Regression Algorithm

```
In [21]: print(linear_reg.score(x_test,y_test))
         0.6914155526647017
```

Test the accuracy value for Random Forest Algorithm

```
In [24]: clf.score(x_test,y_test)

Out[24]: 0.8831565541072046
```

Random forest

Random Forest is a well-known machine learning algorithm that uses the supervised learning method. In machine learning, it can be utilized for both classification and regression issues. It is based on supervised learning, which is the process of combining multiple classifiers to solve a complex problem and improve the model's performance.

"Random Forest is a classifier that contains a number of decision trees on various subsets of a given dataset and takes the average to enhance the predicted accuracy of that dataset," according to the name[5]. Instead than relying on a single decision tree, the random forest gathers the predictions from each tree and predicts the final output based on the majority votes of predictions. The greater number of trees in the forest leads to higher accuracy and the problem of over-fitting is avoided.

Linear Regression

Regression analysis is a type of predictive modelling technique that looks into the relationship between dependent and independent variable. Linear regression is one of the most simple and widely used supervised learning Machine Learning algorithms[3]. It carries out a regression task. Based on independent variables, regression models a target predicted value. Sales, salary, age, product price, and other continuous/real or numeric variables are predicted using linear regression. Simple linear regression is defined as linear regression with only one input variable (x). When there are multiple input variables, this type of linear regression is referred to as multiple linear regressions [4].

Mathematically the relationship can be represented with the help of following equation -

$$Y=a+bX$$

Here,

- Y is the dependent variable we are trying to predict.
- X is the independent variable we are using to make predictions.
- The slope of the line is b.
- a is the intercept [6]

Linear Regression Algorithm Results

```
In [21]: print(linear_reg.score(x_test,y_test))
         0.6914155526647017
```

Random Forest Algorithm Result

```
In [24]: clf.score(x_test,y_test)
Out[24]: 0.8831565541072046
```

OBSERVATIONS

In this paper the two machine learning regression algorithms Linear Regression and Random forest have been compared when trained and tested with the kc_house_data dataset. This has been done in order to study how accurately they, as machine learning methods, predict the prices for the house pricing problem**.**

Table 1 Accuracy Testing Analysis with Classifiers

Accuracy Testing using Machine Learning Algorithms			
Variables	**Algorithm**	**Test Data**	
		Initial State	**Estimated Accuracy**
Primary Train Data	LR	0.22007634	0.691415553
Estimated Train Data	RF	0.11403443	0.882547544

Cross-validation of Linear Regression and Random Forest has proven to be a suitable method to find an acceptable best fitting algorithm for the Model. Random Forest Algorithm is giving very precise Estimation of the house prices. For different Locations it is giving much accurate estimations. Linear Regression gives the least accuracy of 69%. And Random Forest algorithm gives the best accuracy 88%.

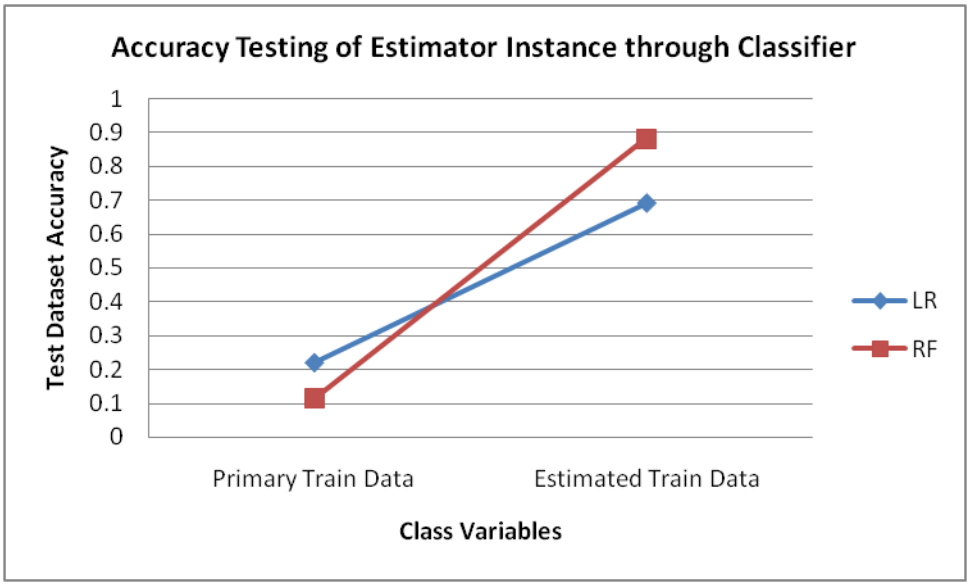

Fig 2 Accuracy Testing of Estimator Instance through Classifier

The accuracy of linear regression algorithm to predict the house price is 0.6914155526647017 and the accuracy of Random Forest algorithm to predict the house price is 0.8825475444503184. The Random Forest was found to consistently perform better than the Linear Regression algorithm with regard to this single variable based prediction analysis.

CONCLUSION

The paper question for this study is to study how well house prices can be predicted by using Linear Regression and Random Forest Regression. In this study we have found that the Random Forest Regression algorithm performs better at predicting house prices than the Linear Regression algorithm. However, there is still a difference between the actual prices in our testing data and the prices predicted by the Random forest regression algorithm.

REFERENCES

1. Nehal N Ghosalkar, Sudhir N Dhage "Real Estate Value Prediction Using Random Forest Regressor," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)
2. Abigail Bola Adetunjia, Oluwatobi Noah Akande*b, Funmilola Alaba Ajala a, Ololade Oyewo a, Yetunde Faith Akande c, Gbenle Oluwadara b "House Price Prediction using Random Forest Machine Learning Technique" The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)
3. https://realpython.com/linear-regression-in-python/
4. https://www.javatpoint.com/linear-regression-in-machine-learning
5. https://www.javatpoint.com/machine-learning-random-forest-algorithm
6. www.stat.yale.edu/Courses1997-98/101/linreg.htm

Cite this article:

S Ponni Jenet, Dr P Revathi, "Accuracy Analysis of House Price Prediction Using Machine Learning Techniques", Journal of Multidimensional Research and Review (JMRR), Vol.3, Iss.1, pp.16-22, 2022