IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 10, October 2020

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

# Bengaluru House Pricing Model Based On Machine-Learning

Drumil Joshi[1], Fawzan Sayed [1], Jai Beri[1]

U.G. Student, Department of Electronics and Telecommunication Engineering, Dwarkadas .J. Sanghvi College of Engineering, Vile Parle, Mumbai, Maharashtra, India[1]

**ABSTRACT**: With the application of data science and machine learning algorithms in Jupyter, the following research displays the potential to predict the market prices of houses in the Bengaluru region based on the following parameters: location, BHK make-up, number of bathrooms, total area in square feet, number of balconies, specific society, nature of real estate (carpet area, built-up, super built-up). Aforementioned software makes use of regression models with 65 and 35 percent training and testing allocation respectively. Models used are linear regression, lasso regression, ridge regression, decision tree, random forest and support vector machine(SVM) to make accurate predictions regarding sale price. The dataset for the same has 15,000 diversified entries and reliability of output values has been verified with RSME and cross-validation reports.

**KEYWORDS**: Data Science, Machine Learning, Jupyter, Regression Model, Price Prediction, Real Estate, RMSE, Cross-Validation

## I. INTRODUCTION

Over the course of the next twenty years, a large percentage of the Indian population under the poverty line will enter the middle class. With more job security and a stable source of income, the immediate desire to upgrade basic amenities like shelter will be observed. As purchasing power rises, the demand for luxury goods rises with it, even if the nature of said item is a necessity for survival. With this in consideration, it is safe to say that the real estate market in India will witness tremendous growth in the future and the development of it is also extremely crucial when it comes to the overall economic development of a region, primarily through house sales, but also through employment opportunities and associated purchases. India's housing prices grew 5.8 % YoY in Mar 2020, following an increase of 6.7 % YoY in the previous quarter, with an average growth rate of 6.6% with respect to the past and the present, and according to a survey in 2011, India stands second in the world for the most number of households(24.67 crores). With this in mind, a problem, region-specific to Bengaluru, is that the number of people trying to purchase a household is enormous. Additionally, the real estate market is directly connected to the banking and finance sector through bonds and loans and real estate is also a popular choice for investments. Therefore, the times demand for a software that is meant for accurate price-prediction while keeping up with the ever-changing and volatile real estate market dynamics. Said software should also be able to address the specific amenities and needs that an individual desires, i.e. a tailor made output according to the exact demands of the customer. The old price-prediction model is based on the comparison of actual house price and estimations. This system is manual and labour-based, not computerized. Such a system is not only extremely inefficient but also gives rise to a major amount of errors. A frequent argument against computerized models is that they lack personalization and a human-touch, therefore, are bound to give unreliable outputs. However, new-age technologies have now evolved to a point where they are able to study thousands of inputs with various parameters, thus compensating for human-touch.

## II. RELATED WORK

The concept of real estate price prediction algorithms was initially made popular by Prof. Murtaza Haider, a civil engineer turned data scientist. His findings and discoveries offered great insight to the Canadian Transport Management and also the housing market in Canada. After realising its immense importance, such algorithms were replicated/innovated around the globe especially in the United States, with companies like Zillow prototyping this model on their website. However, in the last decade with the fast economic development happening, the eastern world, with countries like Malaysia and India, has initiated processes to build such models. All of the famous research projects done on this topic make use of the same set, decisive steps to make their software useful. First, parameters that are most

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

significant to a particular culture/region are selected, such as total sq.ft etc. Next, data from a trustworthy source is analysed manually and digitally to find null data entries and outliers; pieces that do not fit the theme of the research and can cause problems while model-building. Following that, a data set is fed to the model for training and testing, percentage of which varies from project to project. And lastly, the cross-validation of model is done by running varied iterations, and if the nature of output Y(price) matches real data with above 70% accuracy, the users can claim it to be a well-functioning software.

## III. METHODOLOGY

In methodology, an overall layout of the process is described. Various data mining and analysis techniques were used, as shown below (figure 1):

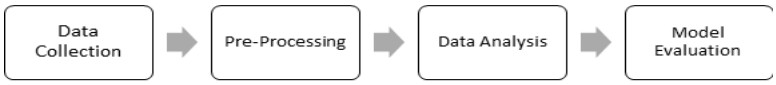


**Figure 1: Project Flow**

### 1) Collection of Data

The open-source dataset was taken from KaggleInc. It consisted of the housing data for the region of Bengaluru. Area type, Availability, Location, Size, Society, Total Square feet, No. of Bathrooms, Balcony, and Price are the parameters mentioned in the dataset. Out of these, Price is the dependent parameter we are trying to predict while the rest are independent parameters. A brief description can be seen in Table 1:

| Parameters | Description | Data Type |
|---|---|---|
| Area_type | States type of area given | String |
| Availability | Date of availability | String |
| Location | Area where located | String |
| Size | Number of BHK | String |
| Society | Type of society | String |
| Total_sqft | Area in square feet | String |
| Bath | Number of bathrooms | Numerical |
| Balcony | Number of balconies | Numerical |
| Price | Price in Lakh Rupees | Numerical |

**Table 1: Parameters**

### 2) Pre-processing

This is wherethe unprocessed data is converted into systematic information void of any inconsistencies. It includes finding lost and unwanted information in the database. The complete database has been tested by NaN and any detection containing NaN will be deleted. Then, columns like Size and Total_sqft are converted to numerical data by using functions as shown in figure 2.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

```
def average_sqft(x):
    y=x.split('-')
    if len(y)==2:
        return(float(y[0])+float(y[1]))/2
    try:
         return float(x)
    except:
        return None
```

**Figure 2: Converting Total_sqft to Numerical values**

Columns like Availability, Bath and Balcony are dropped. Also, Locations contained too many unique values, so values with less than 10 occurrences were all combined into 'Others' location.

### 3) Analysis of Data

We need to find out the characteristic of the dataset to be ableto add any model to our database. Therefore, we need to analyse our database and learn the different parameters and the relationships between the different parameters. We also need to find the outliers present in our database. Outliers are present as some properties (houses) are a bit extreme and can upset the training model.

An approximation was made that for any house, area per bedroom should be at least 300. So, all rows where this condition was not met, were dropped. Then for every location, the mean price per square feet was calculated and all instances where the price per square feet lied outside the standard deviation was removed to remove the extreme cases (figure 3).

```
def outliers_remove(data):
    df_op=pd.DataFrame()
    for key,subdf in data.groupby('location'):
        m=np.mean(subdf.price_sqft)
        st=np.std(subdf.price_sqft)
        temp_df=subdf[(subdf.price_sqft>(m-st))&(subdf.price_sqft<=(m+st))]
        df_op=pd.concat([df_op,temp_df],ignore_index=True)
    return df_op
```

**Figure 3: Removing Square Feet Outliers**

There were also cases where for the same location and same square feet are, houses with a lower bedroom count were priced more than those with a higher bedroom count. Few of these cases are shown in the scatter plot in figures 4.

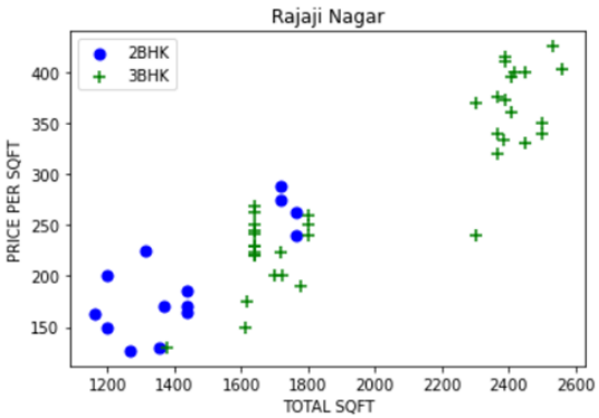


**Figure 4: Errors in Rajaji Nagar**

These cases were also removed using a function (figure 5).

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

```
def outliers_bhk(data):
    exclude=np.array([])
    for location,location_data in data.groupby('location'):
        bhk_stats={}
        for bhk,bhk_data in location_data.groupby('bhk'):
            bhk_stats[bhk]={
                'mean': np.mean(bhk_data.price_sqft),
                'std' : np.std(bhk_data.price_sqft),
                'count': bhk_data.shape[0]
            }
        for bhk,bhk_data in location_data.groupby('bhk'):
            stats=bhk_stats.get(bhk-1)
            if stats and stats['count']>5:
                exclude=np.append(exclude,bhk_data[bhk_data.price_sqft<(stats['mean'])].index.values)
    return data.drop(exclude,axis='index')
```

**Figure 5: Removing BHK Outliers**

Another approximation made was the number of bathrooms can, at most, be only 2 more than the number of bedrooms. So, rows where this was not met is removed. Finally, columns were added corresponding to each unique location and if the house was located in a certain location, the value of that column was set to 1 and 0 for the rest. The Location column was dropped. This was done to provide ease of model training.

We can now see the Correlation Heat Map in figure 6. This shows how much the different parameters are dependent on each other.

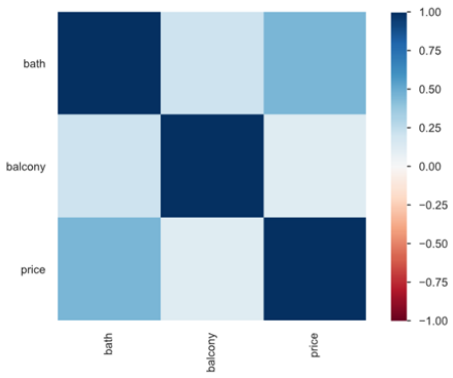


**Figure 6: Correlation Heat Map**

### 4) Application of Machine Learning Models

After the data has been scrubbed for cleanliness and important insights were made with regard to the database, we can implement the appropriate machine learning model for to our database. Hence, six algorithms were selected to predict dependencies it varies from our database. The algorithms which are selected are used as regressors but we train them to predict ongoing prices. The six machine learning algorithms are Linear Regression, Lasso Regression Technique, Ridge Regression Technique, Decision Tree, Random Forest and Support Vector Machine. These are the algorithms will be implemented with the help of Scikit-learn Python Library.Predicted results obtained are plotted to find the distribution of the range in the code operation.

### 1) Linear Regression Algorithm

Regression is a basic mathematical operation including several modelling and analysis techniques. Regression analysis is a tool which explains the relation between the independent variables against its dependencies.Thedependent variable is indicated by the Y, and the independent variables by X. Independent variables as well-known as predictable or descriptive variables. For the most part retrospective, dynamic analysis is considered to be correct ongoing. In simple setup, there is only one independent variations.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

Regression can be expressed as:

□ (□/ □) = □0 + □1□1 + ⋯ + □□□

where β0 is called intercept and βj is called slopes or retraction coefficients. Differences between predicted and the real Y value is called an error (ε) or can be written as □ = □^ - □. After that, the regression equation can be expressed as

The Linear Regression function is displayed as follows:

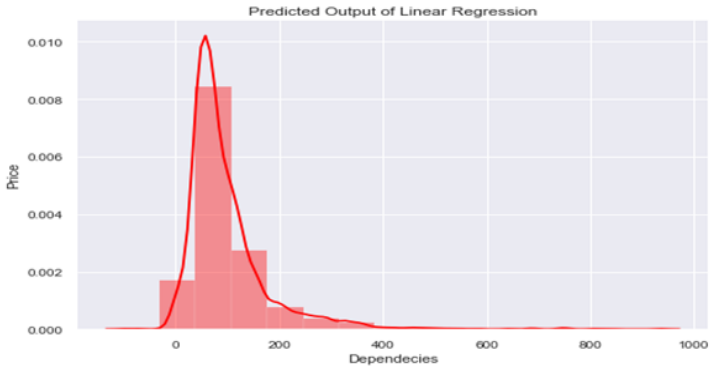


**Figure 7**

### 2) Lasso RegressionAlgorithm

Lasso algorithm is a wonderful way to do regression analysis.Itis used by taxing the magnitude of the feature coefficients and bring the predicted and observed as close as possible to each other. Lasso Regression is known as L1 Regularization Technique. Lasso technique 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 'Remnants Sum of Squares' the square number of errors between predicted values and actual values in the training dataset. α works in partnership that takes various values.

The Lasso Regression function is depicted as follows:

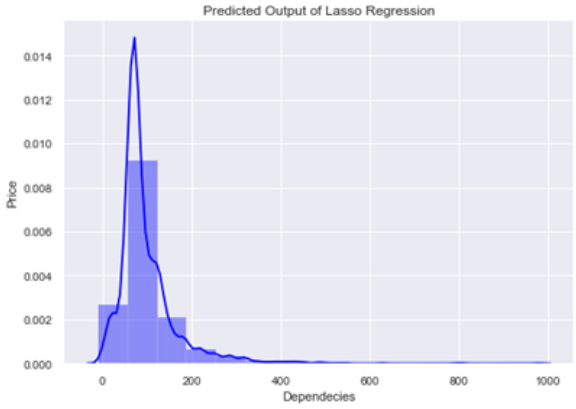


**Figure 8**

### 3) Ridge Regression Algorithm

Ridge regression algorithm is a model tuning method that is used to analyse any data that suffers frommulticollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

The cost function for ridge regression:

$$Min(||Y - X(theta)||^2 + \lambda||theta||^2)$$

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

Lambda is the taxation term. λ given here is denoted by an alpha parameter in the ridge function. Hence, by changing the values of alpha, we are controlling the penalty term. Higher the values of alpha, bigger is the penalty and therefore the magnitude of coefficients is reduced.

The output of Ridge Regression function is as follows:

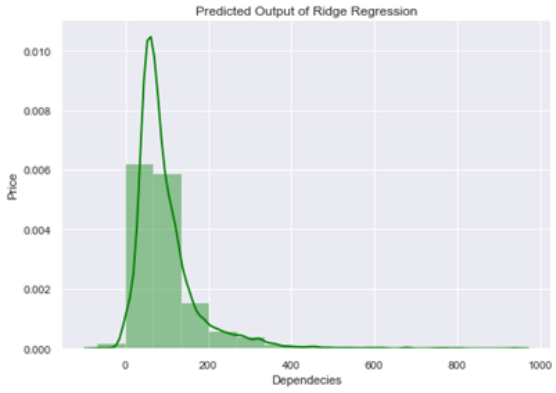


**Figure 9**

### 4) Decision Tree Algorithm

Decision trees are one of the best and efficient advanceda highly monitored algorithm. The model developed is highly efficient in predicting the outcome withgreat score. This model can be used for regression as well as classification.For us we need further predictions on ourtarget price which is why our problem is with the type of regression.We use a binary tree that will do just that re-divide the forecast vector differently subsets such as our target value Y is higher homogeneous. The best part of the model is that it has low variance between parameters. However, one of the most important challenge in decision-tree algorithm is that it can lead to high over fitting model. In the worst-case scenario, it will process the leaf node for each section and thus provide full accuracy.

The Decision Tree Regressor function implemented is as follows:

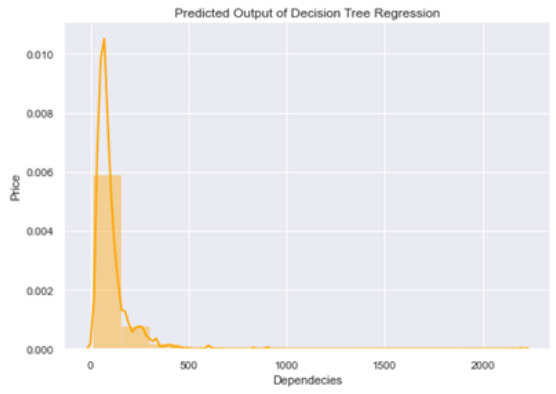


**Figure 10**

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

### 5) Random Forest Algorithm

Random Forests is one of supervised learning algorithm. The model can be used for regression and classification. The forest contains trees.Random Forest is a combination of N number of decision trees. Random forests form decision-making trees in randomly selected data samples, get predictions for each tree and select the best solution by voting. It also provides a good indication of the importance of the feature. Unplanned forests are considered to be the most accurate and powerful method due to the number of decision trees that participate in the model development. It prevents the model from overcrowding.

The Random Forest is implemented as follows:

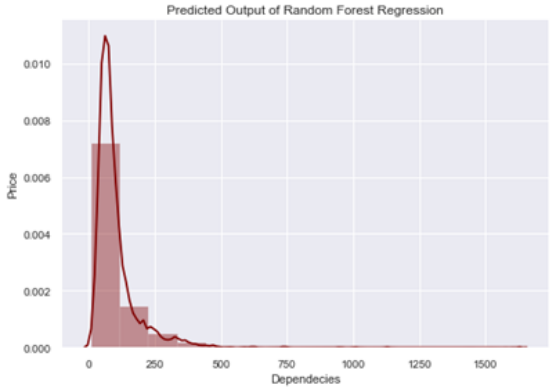


**Figure 11**

### 6) Support Vector Machine

Support Vector Machine (SVM) is a simple Directed Machine Learning Algorithm used for regression and classification. Basically, SVM detects a hyper-plane that creates a boundary between data types. In 2-D space, this hyper-plane is nothing but a line. In SVM, we organize each data object in the database in the N-dimensional space, where N is the number of elements / attributes in the data. Next, we find the appropriate hyperplane to separate the details. So, by this, you should have understood that naturally, SVM can only do binary separation (i.e., choose between two classes). The C value displays the relationship between f and ε.

The SVM algorithm is applied as follows:

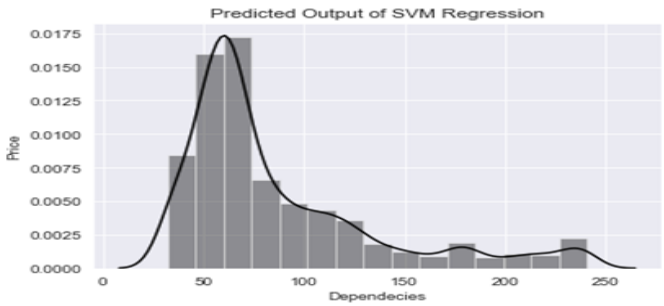


**Figure 12**

## IV. RESULTS

The following table displays the results of various machine learning algorithms that are implemented. We have considered the following points to determine the best machine learning algorithm which are depicted in Table 2.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

| Model Name | Accuracy Score | Cross Validation Mean Score | RMSE Score |
|---|---|---|---|
| Linear Regression | 81.98% | 83.81% | 36.76 |
| Lasso Regression | 66.82% | 69.52% | 49.88 |
| Ridge Regression | 81.61% | 83.53% | 37.13 |
| Decision Tree | 61.17% | 67.77% | 53.99 |
| Random Forest | 69.71% | 78.67% | 47.5 |
| Support Vector Machine | 56.00% | 47.33% | 57.43 |

**Table 2**

From the above table we find that Linear Regression Algorithm gives high accuracy score and after cross validation the score increases and has low RMSE value as compared to other Machine Learning Algorithms

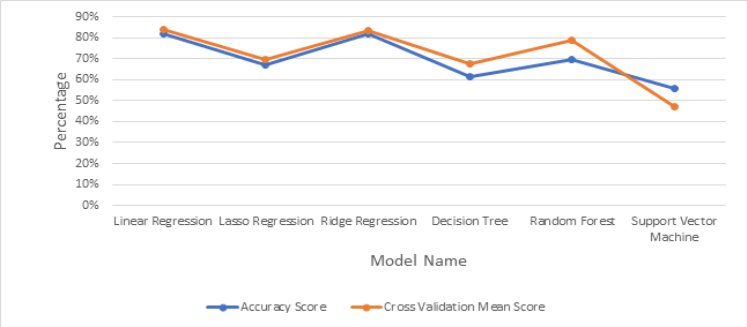


**Figure 13**

The below bar graph shows the comparison of the RMSE values of different Machine Learning Algorithms applied:

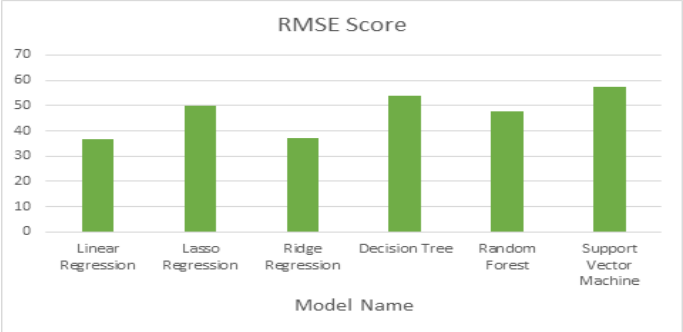


**Figure 14**

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 10, October 2020||

## V. CONCLUSION

In this research paper, we predicted housing prices by implementing ML. We have followed the step-by-step process of studying the database and correlation between features. We can therefore select the features for they are homoscedasticity in nature. These feature sets were then assigned as input to six algorithms and csv file created containing predicted housing prices. So do we calculate the performance of each model used performance metrics are different and compare them based on those metrics. For future work, we recommend that we work on big data which can produce a better and more realistic image about the research. We have implemented only a few Machine learning algorithms. However, we need to train many other regressors and understand how they predict ongoing prices. By deploying this research work, we can make a new estimator for housing prices which is not currently available in India

## REFERENCES

[1] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, "A Hybrid Regression Technique for House Price Prediction", December 2017.
[2] Ayush Varma, Abhijit Sharma, Sagar Doshi, Rohini Nair, "House Price Prediction Using Machine Learning And Neural Networks", INSPEC number 18116205, April 2018 .
[3] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth EmaFebrita, Wayan Firdaus Mahmudy, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017. [4]
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.
[5] Pow, Nissan, Emil Janulewicz, and L. Liu. "Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal." (2014).
[6] Hromada, Eduard. "Mapping of real estate prices using data mining techniques." Procedia Engineering 123 (2015): 233- 240.
[7] J. Gallego, & Mora-Esperanza (2004). ―Artificial intelligence applied to real estate valuation: An example for the appraisal of Madrid‖. Catastro: 255-265.
[8] 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.
[9] 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.
[10] 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

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details