International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Metropolitan City House Price Prediction

Ankit Kumar , Aniket Kumar Singh , Ankit Yadav , Anand Mohan Yadav , Ayush Kumar

Niranjan Kumar (Mentor)

Department of Computer Science & Engineering

Bachelor of Technology

BABU BANARASI DAS INSTITUTE OF TECHNOLOGY & MANAGEMENT-054, LUCKNOW, INDIA

ABSTRACT

In this paper, we are predicting the sale price of the houses using various machine learning algorithms. Housing sales price are determined by numerous factors such as area of the property, location of the house, material used for construction, age of the property, number of bedrooms and garages and so on. This paper uses machine learning algorithms to build the prediction model for houses. Here, machine learning algorithms such as logistic regression and support vector regression, Lasso Regression technique and Decision Tree are employed to build a predictive model. We have considered housing data of 3000 properties.

Predictive models for deciding the sale price of houses in metropolitan cities is still remaining as more challenging and trickier task. The sale price of properties in cities like Bengaluru depends on a variety of interdependent factors. Key factors which may affect the house price include area of the property, location of the property and its amenities. In this system, an attempt has been made to construct a predictive model for evaluating the price based on the factors that affect the price. Modelling study apply some supervised learning techniques such as Bayesian classifier or KNN algorithms. Such models are used to build a predictive model, and to pick the best performing model by performing a comparative analysis on the predictive errors obtained between these models. Here, the attempt is to construct a predictive model for evaluating the price based on factors that affects the price. We build this concept as real time application useful for real estate business and also buyer and sellers.

INTRODUCTION

Real Estate Property is not only the basic need of a man but today it also represents the riches and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Changes in the the real estate price can affect various household investors, bankers, policy makers and many. Investment in real estate sector seems to be an attractive choice for the investments. Thus, predicting the real estate value is an important economic index. India ranks second in the world in number of households according to 2011 census with a number of
24.67 crore. India is also the fastest growing major economy ahead of China with former's growth rate as 7% this year and predicted to be 7.2% in the next year. According to the 2017 version of Emerging Trends in Real Estate Asia Pacific, Mumbai and Bangalore are the top-ranked cities for investment and development. These cities have supplanted Tokyo and Sydney. The house prices of 22 cities out of 26 dropped in the quarter from April to June when compared to the quarter January to March according to National Housing Bank's Residex(residential index). With the introduction of Real Estate Regulation Development Act (RERA) and Benami property Act throughout the country India, more number of investors are attracted to invest into real estate in India. The strengthening and modernizing of the Indian economy has made India as attractive Investment destination. However, past recessions show that real estate prices cannot necessarily grow. Prices of the real estate property are related to the economic conditions of the state [1]. Despite this, we are not having proper standardized ways to measure the real estate property values.

Generally the property values rise with respect to time and its appraised value need to be calculated. This appraised value is required during the sale of property or while applying for the loan and for the marketability of the property. These appraised values are determined by the professional appraisers. However, drawback of this practice is that these appraisers could be biased due to bestowed interests from buyers, sellers or mortgages. Thus, we require an automated prediction model that can help to predict the property values without any bias. This automated model can help the first time buyers and less experienced customers to understand whether the property rates are overrated or underrated. Now, Property prices depend on various parameters in the economy and society. However, previous analyses show that house prices are strongly dependent on the size of the house and its geographical location. We have also considered various intrinsic parameters (such as number of bedrooms, living area and construction material) and also external parameters (such as location, proximity, upcoming projects, etc.) . Then we have applied these parameter values to two different machine learning algorithms. We have considered linear regression model and support vector regression model to predict the price value of the house and compared their output. In this paper, we are predicting house price values using two models i.e. Linear regression, support vector regression, Lasso Regression and Decision Tree with their corresponding accuracy and comparing them based on various error metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), R- Squared value and Root Mean Squared Error(RMSE). In addition to this we will also discuss the significance our approach and the methodologyused.

Literature Review

A. Regression Analysis

i. Hedonic Price Model

The housing market is slightly different from normal good consumption. According to, housing market is unique because it displays the characteristics of resilience, flexibility and spatial fixity. Therefore, hedonic approach is preferred to accurately predict market differential.conceived hedonic model back in 1939, but this research was popularized in the early 1960s with comprehensive uses by Zvi Griliches and Rosen. In the early 1930s, Court used this model to analyse automotive value in pricing and quality characteristics. defined hedonic as "the implicit prices of attributes and are revealed to economic agents from the observed prices of differentiated products and the specific quantities of their characteristics." Following years of progress, Rosen applied the approach to the residential home price study and became commonly included in real estate sector research . Rosen's philosophy or model comprises two separate phases. The regression of a product price on its attributes is performed in the initial stage to calculate the aggregate price of the component. A measure of a goods price will be determined in the first stage, but the inverse demand function cannot be generated at this stage. Thus, the second stage of estimation is needed to identify the inverse demand function that can be derived from the first stage implicit price function. In an earlier study, a study compared three commonly used house price measurement methods which are simple average method, hedonic model, and matching approach. The result found that when adopted on the housing market, two methods that are simple average method and matching approach were proven biased. Thus, the hedonic model provides the highest results relative to those two most commonly encountered versions. Hedonic pricing model is a statistical model that believes the worth of the property is the sum of all its attributes based on hedonic market theory.

ii. Multiple Linear Regression

Regression analysis is a model used to determine the relationship between variables. In order to evaluate the correlation of the variables, the correlation coefficient or regression equation can be used . Multiple regression models can determine which characteristics are the most important to explain the dependent variable. Multiple regression analysis also allows certain price predictions by capturing independent and dependent variable data. In , the power of the multiple regression model can be seen when the value of the relationship between dependent and independent variables is measured. use multiple regression modelling to describe improvements to an independent variable with a dependent variable. This model can be achieved using the house price projection as separate and dependent variables like house prices, house size, property sort, number of bedrooms, and many more. Therefore, the house price is set as a target or dependency variable, while other attributes are set as independent variables to determine the main variables by identifying the correlation coefficient of each attribute.

B. Support Vector Regression

Support vector regression is a predictive model based on SVM, a neural network that usually has three layers, a powerful form of supervised learning. The model is based on a subset of training data. The advantages of support vector regression are that it is capable of processing non-linear results, provides only one possible optimal solution, and able to overcome a small sample learning issues . The potential to produce market predictions in several markets, including real estate, shows that this model can overcome the non-linear regression problems and small sample learning problems. Moreover, as this model did not depend on probability distribution assumptions, and the ability of mapping the input attribute, either linear or non-linear, this model was commonly used at house price modelling. Support vector regression offers huge benefits in so many aspects as this model can avoid over-fitting problems, while ensuring a single optimum solution by minimizing structural risks and empirical risks . In this field of study, support vector regression is used to collect details on neighborhood, structural and locational attributes.

Methodology

Data Preprocessing “Housing Price in Metropolitan Cities” is a dataset containing more than 300, 000 data with 26 variables representing housing prices traded between 2009 and 2021. These variables, which served as features of the dataset, were then used to predict the average price per square meter of each house. The next step was to investigate missing data. Variables with more than 50% missing data would be removed from the dataset. The variable “Day on market” was removed because of 157, 977 missing data. Any observation which had missing values were also removed from the dataset. Below are a few feature engineering processes which were done to cleanse the dataset, Remove attributes indicating the number of kitchens, bathrooms, and drawing rooms due to their ambiguity. Set the number of living rooms (bedrooms were mistranslated to living rooms) in a range from 1 to 4. Add attribute “distance” indicating the distance of the house from the center of Beijing. Replace attribute “constructionTime” with attribute “age” by deducting the year that the house constructed from the current year (2021). Set minimum values for attributes “price” and “area”. Split the attribute “floor” into attributes “floorType” and “floorH..

1) Data Collection

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

2) Data Preprocessing / Cleaning

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

2.1 Normalizing data

The numerical variables of the data set take on a large range of values and depending on the column these ranges can be quite different. In order for this not to introduce bias normalization has been applied, scaling the numbers to the range [0, 1]. The Euclidean norm has been used for normalizing data in this study, although there are other options. The Euclidean norm is denoted ||x||2 and is calculated with the following formula ||x||2 = p x 2 1 + · · · + x 2 n where x1 · · · x n are all of the values of a feature in the data set. Each value of the feature is then normalized by dividing it by the Euclidean norm. [12] [13] [14] Normalization with the Euclidean norm has been used for all features in the data set since all non-numeric features are encoded into numeric as is explained in the following section.

2.2 Encoding categorical data

Many of the variables of the data set are categorical, and take on a limited set of values. One example is the nominal variable ”Street” which represents the type of road access to the property and takes on the values ”Grvl” for gravel and ”Pave” for paved. Such categorical values can not be interpreted by conventional machine learning algorithms without preprocessing them to a numerical format. There are two types of categorical variables in the data set; ordinal and nominal. The difference is that the ordinal variables carry some kind of natural ordering between them. For example, the ”LandSlope” variable, categorizes the slope of the property using one of three values - gentle, moderate or severe - that are intuitively ordered. Labeling is the simple process of pairing a numerical value to each of the ordinal values so that the ordering is preserved. This can be done by simply assigning integer values starting from 1 for the lowest order value, 2 for the next and so on. [15] Figure 3.1 is an example with four arbitrary entries out of the data set and how they might be encoded .

Land Slope	Land Slope

Gtl	1
Mod	2
Sev	3

2.3 Missing Values

Values for entries in the data set that are empty are not useful for the model and thus have been handled in the pre-processing. Fortunately, the data set is fairly complete, containing only a few missing values. These have been processed differently depending on the column. In the nominal and ordinal columns there are a lot of ”NA” values. Pandas interprets this as an empty cell through coercion, however according to the Data Documentation the value ”NA” represents that a feature is not present rather than that it is unknown. For example, the columns ”PoolQC”, which is an ordinal variable describing the quality of the pool, and ”PoolArea” has the value ”NA” for most properties, indicating that there is no pool rather than the information being unknown. Thus, for these columns the value ”NA” is not interpreted as an empty value. However, there are some entries that are empty because of missing values in nine of the ordinal and nominal columns. Each of them except one have between one and four missing values and one of them has 23 missing values.

Finally the column ”MSZoning” needed cleaning. The column represents a nominal variable that describes the general zoning classification of the sale with values like ”RH” for ”Residential High Density” or ”I” for ”Industrial”. For 29 rows the value of the column is abbreviated with ”(all)”, for example ”I (all)”. As it is not specified in the Data Documentation [11] what this represents we’ve chosen to remove those 29 rows that have such a value from the data set. They represent less than 1 % of the data set.

3) Data Analysis

Before applying any model to our dataset, we need to find out characteristics of our dataset. Thus, we need to analyze our dataset and study the different parameters and relationship between these parameters. We can also find out the outliers present in our dataset. Outliers occur due to some kind of experimental errors and they need to be excluded from the dataset. From the analysis we found out that there existsone or two outliers. The general trend for Sale price over different parameters. 'GrLivArea' and 'TotalBsmtSF' seem to be linearly related with 'SalePrice'. The overall quality of the house and Area rises the sale price of the house rises too! However, Overall quality and number of bathrooms are non-correlated and are independent of each other. Total Basement Area and Ground Living Area are correlated to each other. There exists an outlier in all the graphs of Total Basement Area. This outlier could be present due to experimental errors and hence that observation can be avoided.

4) **Error metrics**

For measuring how good predictions the model makes, four error metrics have been used. Mean absolute error (MAE), Mean squared error (MSE), Median absolute error (MedAE) and Coefficient of determination (R2). They are all defined below.

4.1 Mean absolute error (MAE)

Mean absolute error measures the prediction error by taking the mean of all absolute values of all errors, that is: MAE = Pn i=0 |yi − yˆi | n Where n is the number of samples, y are the target values and yˆ are the predicted values. A MAE closer to 0 means that the model predicts with lower error and that the prediction is better the closer the MAE is to 0. [20]

4.2 Mean squared error (MSE)

Mean squared error is similar to MAE, but the impact of a term is quadratically proportional to its size. It measures the prediction error by taking the mean of all squared absolute values of all errors.

4.3 Median absolute error (MedAE)

The median absolute error (MedAE) is the median of all absolute differences between the predicted value and the target value. In difference to MAE and MSE, the median absolute error is more robust to outliers by virtue of using the median instead of the mean.

$$MedAE = median(|y1 - \hat{y}1|...|yn - \hat{y}n|)$$

A low MedAE means little error and a good prediction. [22] [23]

5)Web Development

Backend:

We will use a Flask server as our backend to host the application locally. In the server folder we will set up two files:

- server.py

The server.py file will be responsible for handling the routes for fetching the location names and predicting the house

price. It also gets the form data from the front end and feeds it to the util.py.

- util.py

The util.py file is the main brains behind the back end. It has a function to load the JSON and pickle file. This file takes the form data from server.py and uses the model to predict the estimated price of the property.

Frontend:

The front end is made up of simple HTML, CSS and JavaScript. The user can select the number of square feet area, BHK, bathrooms and location in the form and hit on 'ESTIMATE' button to get the estimated price. The JavaScript file is responsible for interacting with both the backend flask server routes and the frontend HTML. It gets the form data filled by the user and calls the function that uses the prediction model and renders the estimated price in lakhs rupees (1 lakh = 100000).

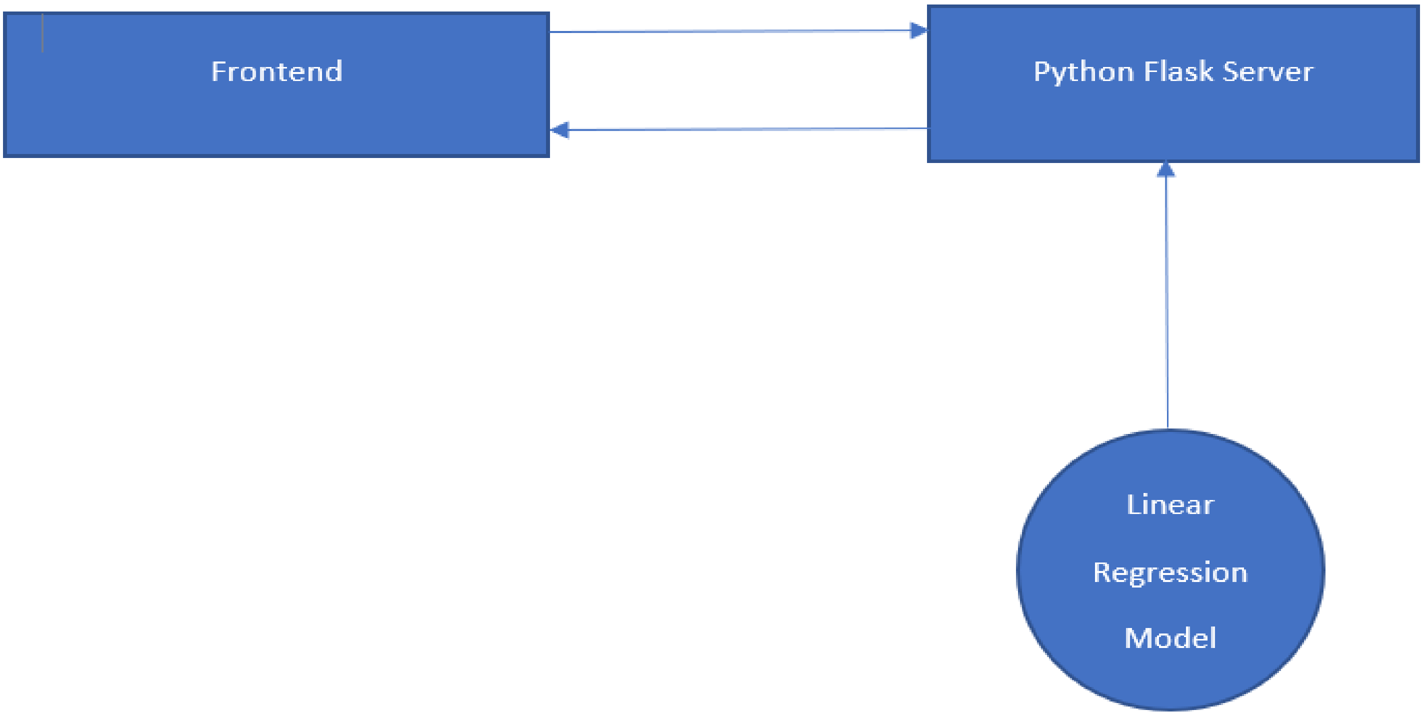
Input Output

Input:

1. Area type-describes the area.
2. Size- in BHK or Bedroom (1-10 or more).
3. Total Square Feet - size of the property in square feet
4. Bathrooms -Number of bathrooms.
5. Balcony- Number of balconies.
6. Road access – Paved road or Gravel road.
7. General shape – Regular or Irregular planning.
8. Flatness – (Banked)- Quick and Significant rise from street to building.(Hillside)- Significant slope from side to side.
9. Slope –Gentle, Moderate or Severe Slope.

Output:

House price prediction using ML supervised learning technique


Frontend
Python Flask Server
Linear
Regression
Model


Bengaluru House Price Prediction
Area in Sq. ft
1000
BHK
Bathroom
Location
1st block jayanagar
ESTIMATE

Progress Work Detail

Currently we have trained a machine learning model for Bangalore city and Main page of Website. Predict house prices in Bangalore using some fairly simple techniques and a bit of feature engineering, on which I would appreciate your feedback. work on a dataset which consists information about the location of the house, price and other aspects such as square feet etc. When we work on these set of data, we need to see which column is important for us and which is not. Our main aim today is to make a model which can give us a good prediction on the price of the house based on other variables.

We are going to use Every ML model for this dataset and see if it gives us a good accuracy or not. From time to time I will also share my understanding of price drivers, outliers and limitations which arise due the absence of certain key features. In website we made home page and all navigations bar Yet.

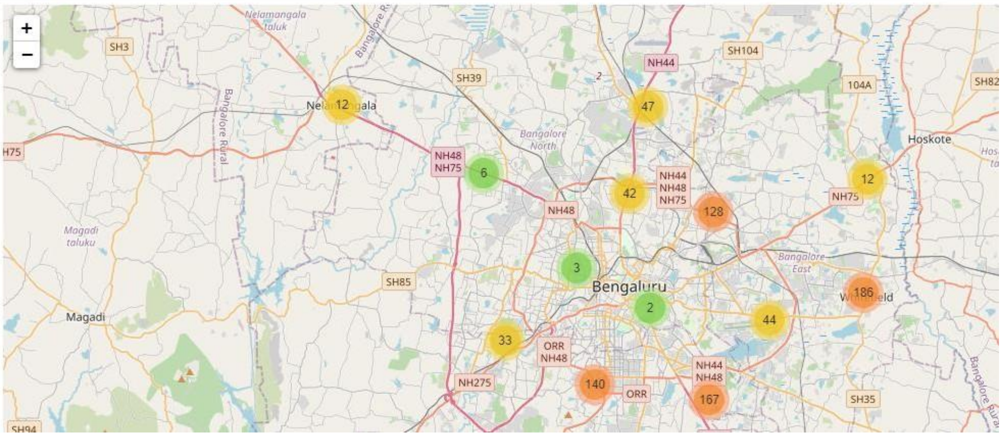
```
In [5]: df = pd.read_csv('Bangalore_new.csv')

In [6]: df.head()
Out[6]:
```

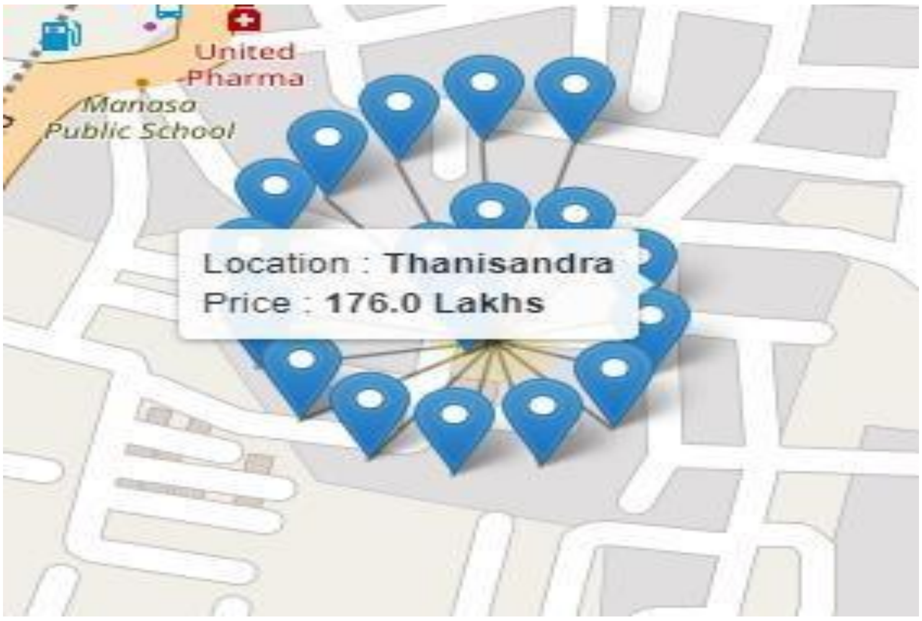
	Unnamed: 0	Price	Area	Location	No. of Bedrooms	Resale	MaintenanceStaff	Gymnasium	SwimmingPool	LandscapedGardens	...	VaastuCompliant	Microw
0	0	300.00	3340	JP Nagar Phase 1	4	0	1.0	1.0	1.0	1.0	...	0.0	
1	1	78.88	1045	Dasarahalli on Tumkur Road	2	0	0.0	1.0	1.0	1.0	...	1.0	
2	2	48.66	1179	Kannur on Thanisandra Main Road	2	0	0.0	1.0	1.0	1.0	...	0.0	
3	3	83.58	1675	Doddanekundi	3	0	0.0	0.0	0.0	0.0	...	0.0	
4	4	68.45	1670	Kengeri	3	0	1.0	1.0	1.0	1.0	...	0.0	

5 rows × 43 columns

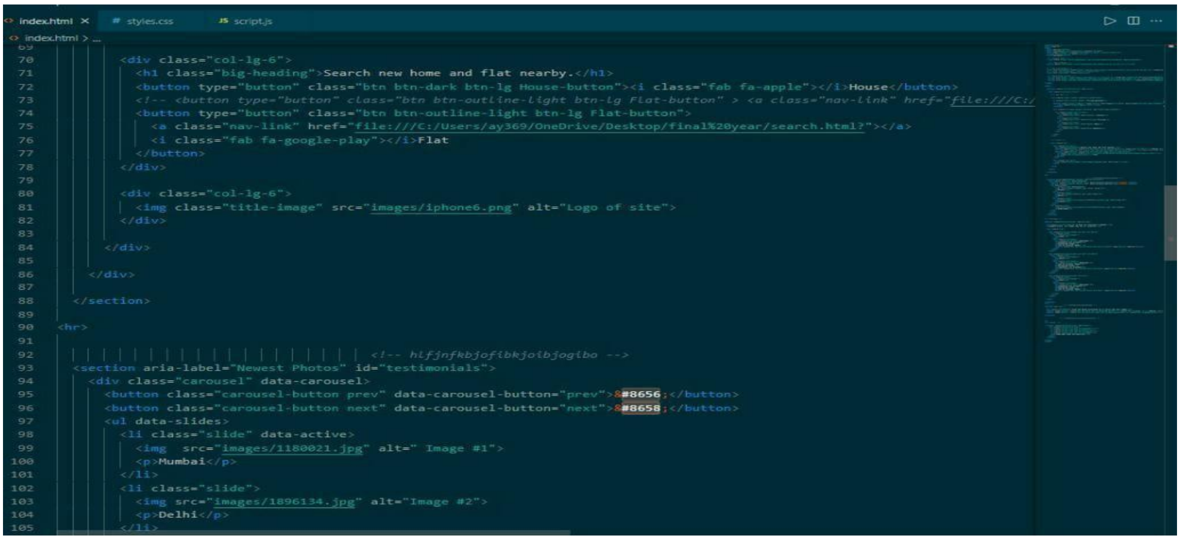
In [21]:
```
from folium import Choropleth, Circle, Marker
from folium.plugins import HeatMap, MarkerCluster
import math

city_map = folium.Map(location=[12.97,77.59], zoom_start=12, tiles='openstreetmap')
mc = MarkerCluster()
for idx, row in df.iterrows():
    if not math.isnan(row['Longitude']) and not math.isnan(row['Latitude']):
        popup = """
        Location : <b>%s</b><br>
        Price : <b>%s Lakhs</b><br>
        """ % (row['Location'], row['Price'])
        mc.add_child(Marker([row['Latitude'], row['Longitude']],tooltip=popup))
    city_map.add_child(mc)
city_map
```

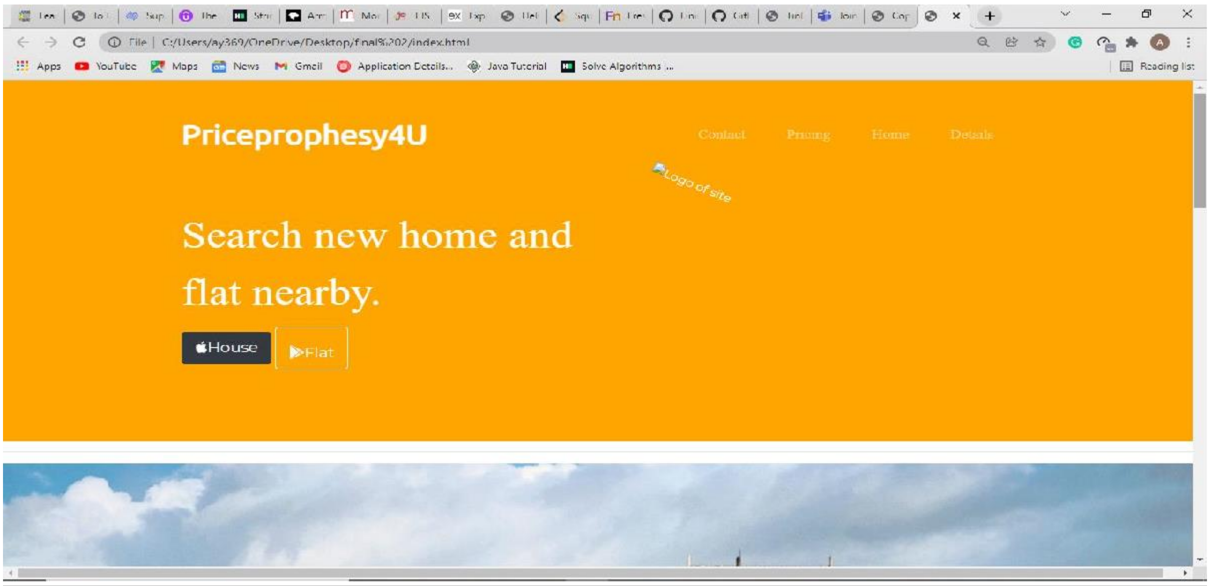
Out[21]:





Website




Priceprophesy4U
Search new home and
flat nearby.
House
Flat

RESULTS

By following the method stated in the previous section, the following results have been obtained. To begin with, the two methods have been tested with different values for the selected hyperparameters as described in the previous section. The results of those tests are presented in the graphs and tables in section 4.2 and 4.3. The values of the parameters for each method that yields the least prediction error has then been used to compare the two methods and the results of that comparison are presented in section 4.4. All tests are performed on the cleaned data, described in section 3.1.

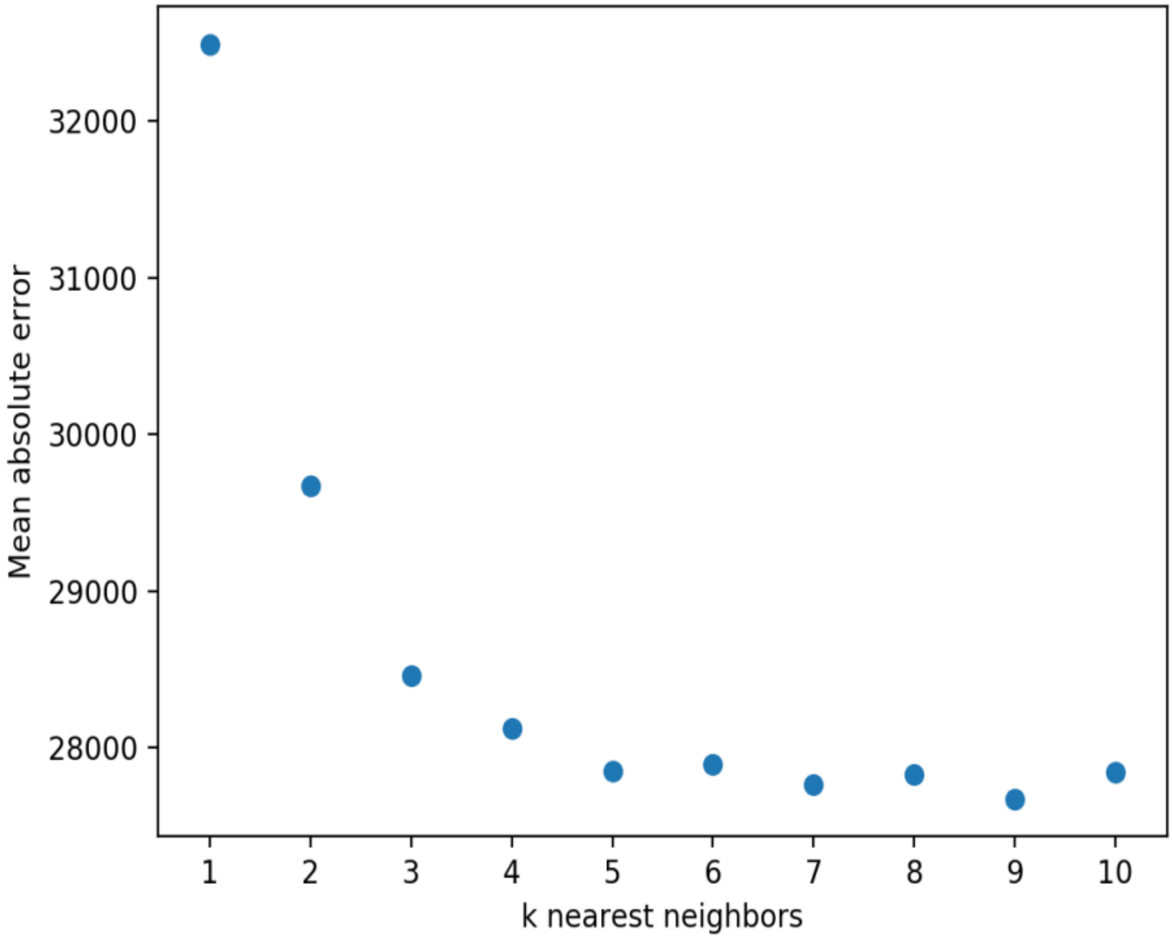
Grid search over hyperparameters

The grid search algorithm was used to find the best set of values for the selected hyperparameters of each algorithm, presented in this section. In short, the following was found: For the k-Nearest neighbour algorithm the best value for the ’k’ hyperparameter is 9 neighbors and for the ’weights’ hyperparameter the best value was ’distance’ (inversely proportional). For the Random forest algorithm the best value for the hyperparameter ’n_estimators’ was 41, for ’max_features’ 63 and for ’criterion’ the best option was ’mse’.

k-NN hyperparameters

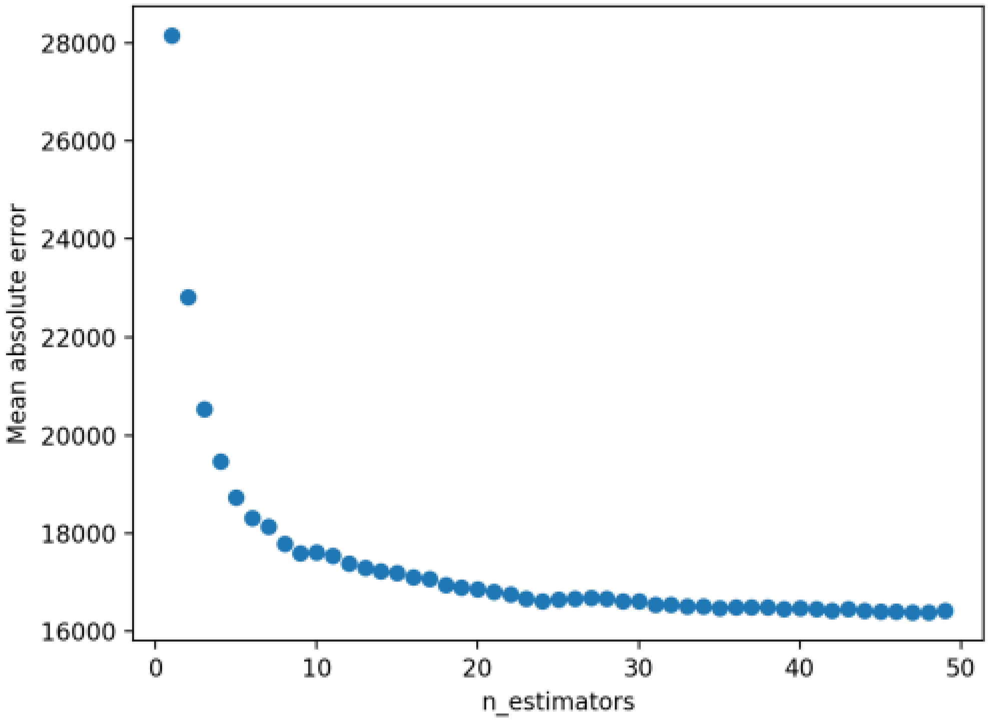
Two hyperparameters have been tested for k-NN with different values. The value of the k number of neighbours to use for the algorithm has been tested for the values 1 through 10 comparing by MAE, displayed in figure 4.1. As can be seen in the figure, the MAE is least when k is equal to 9 which provides the best prediction with the current setup for the k-NN algorithm when k is ranging between 1 and 10. The value for the ’weights’ hyperparameter was set to ’distance’ when testing the different values for the k parameter since this value produced the lowest error in the grid search over the hyperparameters.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930



Random forest hyperparameters

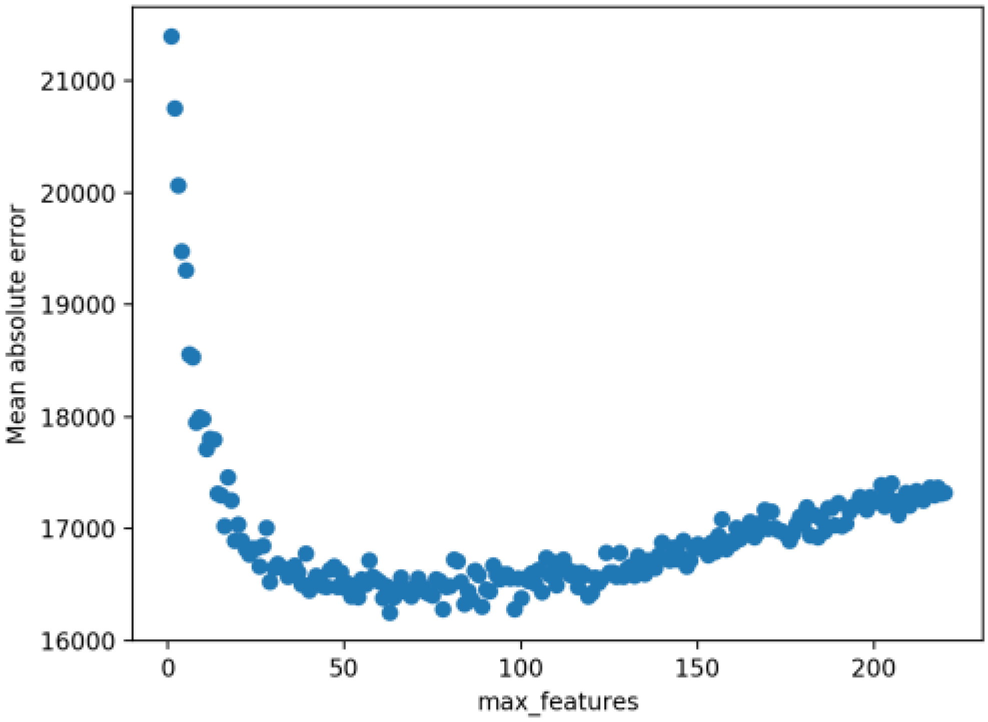
The random forest algorithm has been tested with three different hyperparameters; ’n_estimators’, ’max_features’ and ’criterion’. ’n_estimators’, which represents the number of decision trees in the random forest, has been tested with values ranging between 1 and 50 and the results are presented in figure 4.2 where the values are compared by MAE. When testing the ’n_estimators’ parameter, the ’criterion’ hyperparameter was set to ’mse’ and the ’max_features’ hyperparameter was set to 63 since those values performed best in the grid search.



MAE for Random forests with n_estimators from 1 to 50.

Secondly, the ’max_features’ hyperparameter was tested with values from 1 to 221 which is the total number of features in the cleaned data set (including one-hot encoded features). The results are presented in figure 4.2 where values of ’max_features’ are compared by MAE. The value with the least MAE is 63.

The third hyperparameter tested for Random forest is the ’criterion’ hyperparameter which determines what error measurement to use to measure the quality of a split in the trees. The values tested are ’mae’ which uses mean absolute error and ’mse’ which uses mean squared error. The hyperparameter ’n_estimators’ was set to 41 when performing the tests and ’max_features’ was set to 63, values determined from the grid search.



Algorithm comparison

Finally, errors of the k-Nearest neighbour algorithm and the Random forest algorithm are compared, which gets to the focus of this study. Presented in table are the errors for the respective algorithms running on their selected optimal hyperparameters. On all four error metrics investigated, it is clear that the Random forest algorithm performs considerably better than the k-Nearest neighbour algorithm, with regards to predicting with the smallest error.

Algorithm	Hyperparameters	MAE	R^2	MedAE	MSE
k-Nearest neighbours	k=9, weights=distance	27670.7	0.695988	17663.9	1896058151.2
Random forest	n_estimators=41, max_features=63, criterion=mse	16208.5	0.877811	10135.9	754362031.6

Table Errors of the two methods

Conclusion & Future Work

6.1 Conclusion

In this research paper, we have used machine learning algorithms to predict the house prices. We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the parameters. Thus we can select the parameters which are not correlated to each other and are independent in nature. These feature set were then given as an input to four algorithms and a csv file was generated consisting of predicted house prices. Hence we calculated the performance of each model using different performance metrics and compared them based on these metrics. We found that Decision Tree overfits our dataset and gives the highest accuracy of 84.64%. Lasso gives the least accuracy of 60.32%. Logistic Regression and Support Vector Regression giving an accuracy of 72.81% and 67.81% respectively Thus we conclude that we implemented classifiers to the problem of regression to check how well can classifier fit to regression problem . For future work, we recommend that working on large dataset would yield a better and real picture about the model. We have undertaken only few Machine Learning algorithms that are actually classifiers but we need to train many other classifiers and understand their predicting behavior for continuous values too. By improving the error values this research work can be useful for development of applications for various respective cities.

6.2 Future Work

There are quite a few things that can be polished or add in the future work. though, we were able to identify most of the residential areas. There may be some more places that have housing complexes or multi-story apartments which are located in commercial areas. Such apartments were not included in this paper and can be counted in future to give a more accurate result. With more and more demand for housing in metropolitan cities, there is a definite increase in the number private builders that provide real estate with additional amenities to attract more customers.There are several other models available that can be implemented for prediction. Data given as input to such model should be compatible with the tool used and the operators involved in the process. Also, more number of data sets can be used to increase the accuracy of the model. The main objective of using a different model should be to reduce the calculation time and carry out the whole process in ease

REFERENCES

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

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

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

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

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

[6] 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.

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

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

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

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

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

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

[13] https://www.kaggle.com/ohmets/feature-selection-forregression/data.