Real Estate Classification And Price Prediction Website

Karan Nipurte[1], Suraj Nayak[2], Manish Bhagat[3], Vrushali Bhamare[4]

[123]*B.E. Student, Department of Computer Engineering, Shivajirao S. Jondhale College of Engineering, Dombivli East, Thane, Maharashtra 421201, India.*

[4]Asst. Professor, *Department of Computer Engineering, Shivajirao S. Jondhale College of Engineering, Dombivli East, Thane, Maharashtra 421201, India.*

---***---

Abstract -*We are presenting a website which will help the user or citizen to find the perfect residential area or place as per required specification and desired location. It is generally acknowledged that the prices of real estatearehighly complicated and are interrelated with a multitude of factors. It will be advantageous if the parties to have some insights to some degree. This platform may also help the broker or dealer to keep track on the properties that need to be sold. We are further implementing machine learning algorithms and data cleaning process for easy retrievalof data required in machine learning model.A case study was carried out on housing price determinants of a sample project using this model. The results concerning the efficiency of the proposed framework in terms of accuracy and computational time are also presented. It shows that more accurateprice prediction of real estate can be acquired with the linear regression model.*

***Key Words*: Linear regression model, Outlier detection, Data Cleaning, Flask Server, Feature Engineering.**

1.INTRODUCTION

Real Estate Industry is both capital-intensive, highly related industries and industries essential to provide the daily necessities. However, the real estate pricing models and methods of research rarely receives the critical attention and development it deserves. As the real estate projects more heterogeneous, it is not possible adopted a uniform pricing model and methods as for other products. As a result, the real estate pricing model and is relatively backward. This paper considers the determinants for housing price by using the integrated linear regression machine learning algorithm. The present work intends to integrate linear regression with python flask to determine properly the weights of neural network, making up for the defects of BP algorithm. Multiple linear regression is a useful approach to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data.There are two main advantage of using multiple linear regression model to analyze data. First, multiple linear regression has the ability to determine the relative influence ofone or more predictor variables to the criterion value. Second, it has the ability to identify outliers or anomalies. In the present paper, a pricing model of real estate project was taken as the subject investigated. Based on data of Bangalore city, the prediction model of pricing was developed by linear regression machine learning algorithm, providing theoretical guide for real estate project and tools design.

1.1 Problem Definition

To develop an efficient system which will locate the available property as per the user request as much fast as possible and to develop a website which will be able to detect ser location and can suggest the best available property option to him/her.

Using dataset(provided by the survey) of project requirement skills by making use of linear regression model algorithms and machine learning algorithm to build interface, which is user friendly and easy to use. The interface should have following key properties:(1) Simplicity: the pre-processing and the search are very simple, and only bitwise logical are used for locating area. (2) Real time: The user should get the exact time when the property place is vacant for further plans. (3) No buffering: The text does not need to be stored for its time to time retrieval.

2. Literature Survey

By referring the papers which have already worked in this area we got to know the different task that are needed in Implementation of machine learning model, Data cleaning, Outlier detection, feature engineering etc.

In the paper titled "Fusing Neural Networks, Genetic Algorithms and Fuzzy Logic for Analysis of Real Estate Price", author Huawang Shi used ANN for implementing the real estate project, he further stated that, It is generally acknowledged that the price of real estate was highly complicated and was interrelated with a multitude of factors. It will be advantageous if the parties to a dispute have some insights to some degree. Their paper introduces a hybrid genetic algorithm (HGA) approach to instance selection in artificial neural networks (ANNs) for the price of real estate. From above paper we got the main idea about our system and the website. [2] Then we have paper titled "Weather Analysis to Predict Rice Cultivation Time Using Multiple Linear Regression to Escalate Farmer's Exchange Rate" by the authors Luminto&Harlili, they have detailed instructions on applying the linear regression model.Here weather analysis is conducted by retrieving weather data from National Weather Forecast and Farmer's Exchange Rate data from National Statistics Authority for the past 1 year and using the obtained data to build a regression model using Multiple Linear Regression (MLR) to determine the correlation between weather and FER. From above paper we have got idea about our main model wiz Linear regression model. [3]The paper titled "E-Clean: A Data Cleaning Framework for Patient Data" by authors HasimahHj Mohamed, Tee Leong Kheng, Chee Collin and Ong Siong Lee gives many techniques for cleaning the data. Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data cleaning system are needed to support any changes in the structure, representation or content of data. There are three parts in the cleaning process, i.e. extract the invalid value, matching attributes with valid values and data cleaning algorithm. This Besides that, parsing techniques is also use for the identification of dirty data. This cleaning system isvery useful system to clean off those inconsistencies, duplicates and anomalies data in the database. [1] The paper titled "An Outlier Detection Algorithm Based on the Degree of Sharpness and Its Applications on Traffic Big Data Pre-processing" by authors Zhonghao Wang and Xiyang Huang gave the idea about detection and removal. In this paper, a new outlier detection algorithm is proposed, which combines the image processing method with the data processing method to detect the outliers effectively. In this algorithm, a measure in image processing, degree of sharpness, is adopted to detect the outliers at the first time. The proposed algorithm can be easily applied on the applications of data pre-processing, equipment fault diagnosis, credit fraud detection, traffic incident detection etc. The proposed algorithm is a non-statistical learning method. Compared to the classical outlier detection methods with statistical learning, it has no iterative processes. Thus, it can detect the outliers with lower time cost. In the following research, we will take measures to improve the outlier detection rate further. [5]

3. DESIGN AND IMPLEMENTATION

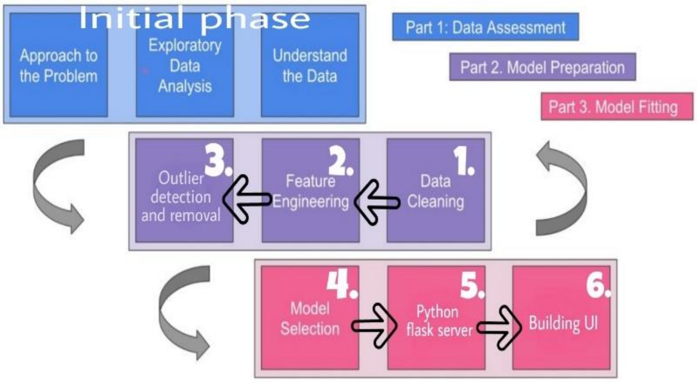


Fig-1: Flow design

IMPLEMENTATION:

1.DATA CLEANING:

(1) We downloaded dataset into pandas to perform data cleaning. (2) We got dataset for particular city through Kaggle.com for particular city on which we have to perform operations. (3) We grouped entities by column by groupby function. (4) To keep model simple for required retrieval purpose we are going to neglect (drop) some columns from available dataset like availability, society, balcony n the dataset. (5) Then we performed the data cleaning process, for that we first perform isnull().sum() function to check the NA values, then we drop those rows by using dropna(). (6) For size column we another unique BHK column, so we used lambda function. (7) Then we rectified some error present in dataset like wrong square feet or mean of range of the square feet by applying some functions.

2. FEATURE ENGINEERING:

(1) For outlier detection and removal in later stage we found out price per sq.mt (by dividing sq.mt and price column) dataframe. (2) Then we found out unique location, There are so many (i.e. 1304 locations), so we came across dimensionality curse, to came with solution we had various techniques like 'Other' category. (3)We grouped as many 1000's of location and got unique 242 locations.

3. OUTLIER DETECTION AND REMOVAL:

(1) Outlier are datapoints which are data errors or extreme dataset or of domain knowledge. (2) Using some threshold value we got genuine data for dataset required (i.e. by using sq.ft per BHK). (3) Then we had to remove extreme values for price per square feet as it is unusual form of data, so we filtered out point which are beyond standard deviation. (4) Then we have a function to check behaviour of dataset like 2 BHK home prize is higher than the 3 BHK homes, so we frame the behaviour (It draws the scatter plot)

i.e. for 'Rajaji Nagar'

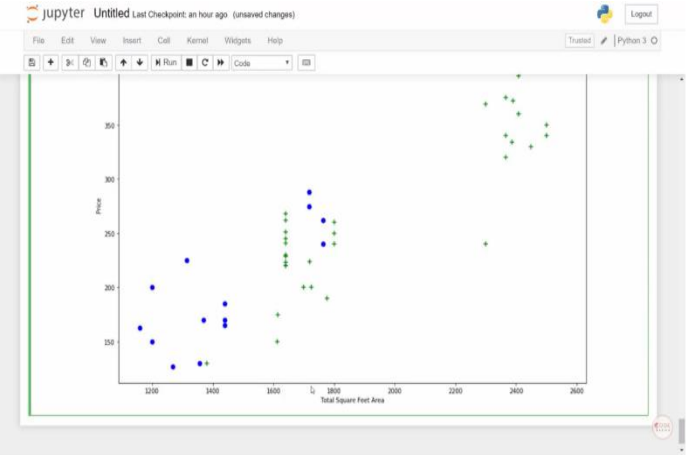


Fig-2: Scatter Plot (Rajaji Nagar)

(6) Plot histogram to see how many apartments in particular area per square feet.

Between 0 to 10,000 Rs per sq.ft we have most datapoints.

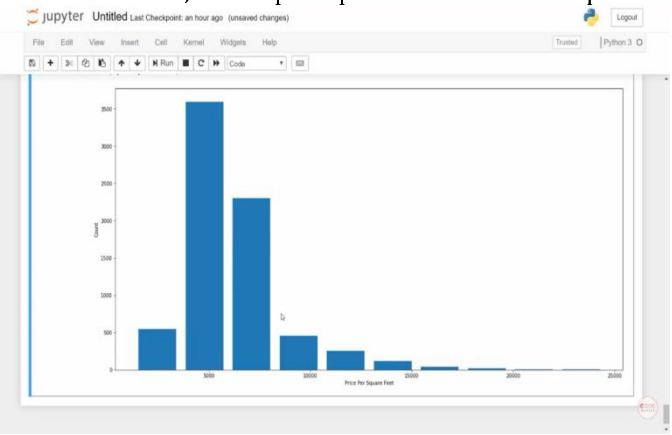


Fig-3:Histogram (Price per square feet)

(7) We removed datapoint where Bathroom number are greater than BHK by 2(i.e. Bath > BHK+2). (8) At the end we created new dataframe to drop feature like price per square feet and size as they are not necessary to implement machine learning model.

4. MACHINE LEARNING MODEL:

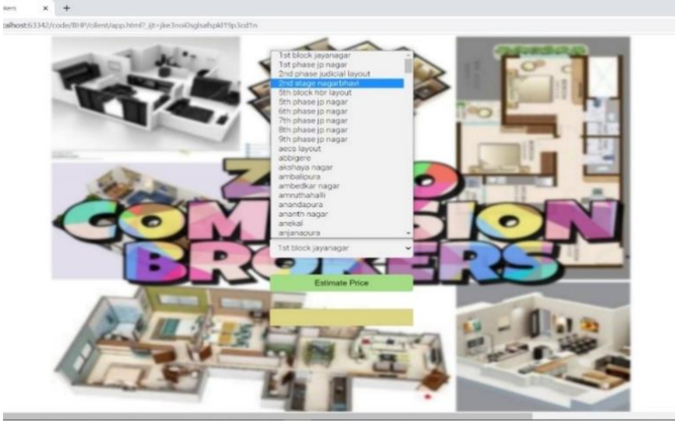
(1) In this module we built a machine learning model and then we used K-fold and Grid Search CV for best algorithm. (2) We convert text to numeric values as in machine learning algorithm it does not allow text but numeric, we did it using one hot encoding (Dummies). (3) We divide dataset into train and test dataset, so we imported train set selection method from sklearn model selection. We got 20% dataset for test sample and other for training. (4) We created linear regression model (using fit and score we got score of 0.845 for it). (5) For K-fold cross validation we imported needed methods and created shuffle split for cross validation. It will randomize our sample, so each fold will have equal distribution of the dataset samples. (we got scores 0.82, 0.77, 0.85, 0.80, 0.83). (6) With grid search CV we checked for Lasso and decision tree regression, it gives best scores and parameters and scores for the run, lasso gave 0.68 score, decision tree gave 0.72 score, linear regression model give 0.81 score this time. (7) So we concluded earlier linear regression model with 0.845 score was the best amongst all for our dataset, so we created predict prize function for our dataset which accepts location, sq.ft, bath, BHK. We also used same function to check the result i.e. predict value(). (8) Now, we exported our model to pickle file so that python flask server can access data for the website. (9) Import pickle file then passed model (pass classifier as argument), When executed it will export file. (we also need column information in json file).

5. PYTHON FLASK SERVER:

(1) It will serve HTTP request made from UI and predict home prices. (2) In this stage we are going to work in server folder of the main file. (3) In this we imported flask module in which we write python service to make HTTP request. (used anaconda for flask). (4) We used app.run() in main function of application on specific port exposed HTTP end points with app.route(). (5) First routine to return location of Bangalore city, created subdirectory in server and copied art effect client to server (until file contain core routine, run same routine to check location names).(6) Created predict function in util file for getting price using two dimensional array. (7) Another routine is created to predict the home prize which take HTTP post method. (8) Used postman application to test HTTP call like post and request (predict home prices).

6. BUILDING UI (WEBSITE):

(1) In client folder we built a HTML, CSS, Javascript based website by visual code.(2) We used HTML for structure of the UI and anatomy, CSS (cascaded style sheet) contains color, look and feel of the website, javascript contains dynamic code which make HTTP call for the backend. (3) According to our layout we made structure fields in HTML. (4) Then we organized the buttons at the backend and added our website edited background, written two more functions to get Bath and BHK value by switchfield and then made the post call (jquerry). (5) In title field we named our website as ‘Zero Commission Brokers’.

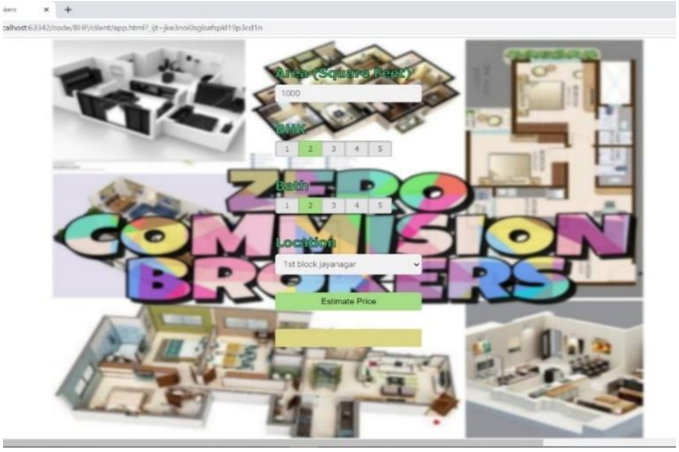


4. Results

MAIN WINDOW:

Fig-4: User Interface

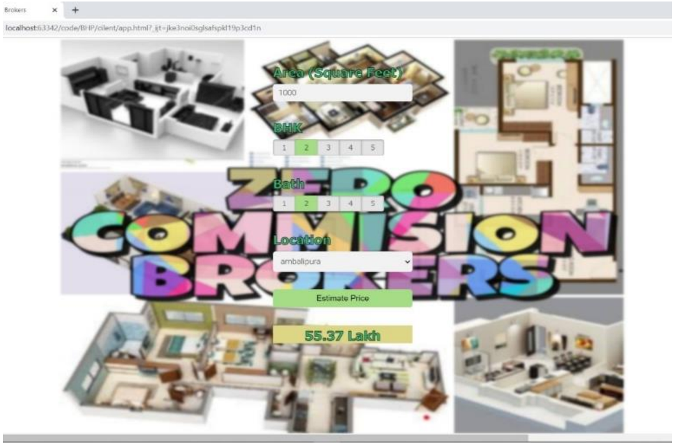
ENTER FIELDS OF INFORMATION AS PER REQUIRED:



Area (Square Feet): 1000
BHK: 1 [✓ 2] 3 4 5
Bath: 1 [✓ 2] 3 4 5
Location: 1st block jayanagar
Estimate Price

Fig-5: Entering UI fields

FINAL RESULT AS PER REQUIREMENTS OF USER:



Area (Square Feet): 1000
BHK: 1 [✓ 2] 3 4 5
Bath: 1 [✓ 2] 3 4 5
Location: ambalipura
Estimate Price
55.37 Lakh

Fig-6: Final Result Window

5. CONCLUSION

It is generally acknowledged that the price of real estate was highly complicated and was interrelated with a multitude of factors. It will advantageous if the parties can get some insight of their desired property to some degree. Using train dataset in our website, The linear regression model can easily predict nearby exact price in less amount of time. We have concluded that Linear regression model is the best out of all available machine learning model by calculating their score to reach the result for our type of dataset. We can increase accuracy of result and decrease the time required for estimating price by providing more percentage of data for training of model and by providing more correct data values so the time required for data cleaning and outlier detection will be less. The main motto of our project is to build a system for the people who are aged and disabled, with our easy UI and easy model one can easily find required available property for investment. The system can also be modified and implemented in government areas for monitoring the properties which are unauthorized and keep track on them.

ACKNOWLEDGEMENT

We are very thankful to our project guide Prof. Vrushali Bhamare, for her valuable support and constant guidance. In spite of her extremely busy schedule, she never failed to take time out for us to help solve our problems and clear our doubts. We would like to take this opportunity to thank our Head of Department, Prof. P. R. Rodge for his motivation, guidance and support. We would also like to thank Dr. J. W. Bakal, Principal, SSJCOE, Dombivli, for his invaluable support and for providing an outstanding academic environment. We acknowledge all the faculties of the Computer Engineering Department for their advice during various phases of this project work.

REFERENCES

[1] HasimahHj Mohamed & Tee Leong Kheng, "E-Clean: A Data Cleaning Framework for Patient Data", 2019.

[2] Huawang Shi, School of Civil Engineering, Hebei University of Engineering, "Fusing Neural Networks, Genetic Algorithms and Fuzzy Logic for Analysis of Real Estate Price", 2018.

[3] Luminto & Harlili, School of Electrical Engineering and Informatics Institute technology Bandung, "Weather Analysis to Predict Rice Cultivation Time Using Multiple Linear Regression to Escalate Farmer's Exchange Rate", 2016.

[4] Zhihong Zhou and Jiao Mo School of Science Beijing University of Posts and Telecommunications Beijing, "Data Imputation and Dimensionality Reduction Using Deep Learning in Industrial Data", 2018.

[5] Zhonghao Wang, Xiyang Huang, School of Optical-Electrical and Computer Engineering University of Shanghai for Science and Technology, "An Outlier Detection Algorithm Based on the Degree of Sharpness and Its Applications on Traffic Big Data Pre-processing", 2019

[6] www.w3schools.com/w3css/default.asp (For text fonts, background, text font and formats, button and switchfield format in our UI)