HOUSE PRICE PREDICTION USING VARIOUS REGRESSION TECHNIQUES

Arpit Bajpai\textsuperscript{1}, Aman Chauhan\textsuperscript{2}, Pawan kumar\textsuperscript{3}

\textsuperscript{1,2,3} School of Computer Science and Technology, Galgotias University Greater Noida, India
\textsuperscript{1}bajpaiarpit159@gmail.com, \textsuperscript{2}devachauhan@gmail.com, \textsuperscript{3}Pk460998@gmail.com

ABSTRACT:

House price forecasting is a key aspect for any real state business. The literature shows us some great knowledge from historical data of markets dealing against property. Here, we applied some of the machine learning techniques to track the previous historical property transactions and to discover a useful model for house buyers and sellers as well. We observe a large set of discrepancy in house prices which does not sounds good. Earlier the predictions were made manually and prices were being calculated via pen and paper. We came up with the idea to remove this bug and prices would be calculated by a machine learning model. This model would be taking various input parameters about the location and based on that it will predict the prices. To implement this we would be using regression techniques and would be training our model on the various different dataset. This model would prove to be a great one for both real estate and customers. It will help in maintaining the transparency between both of them. In order to make this model robust we would train this on the large and wide set of dataset so that it produces approximate results.

Keywords: Rodent-Hole Mining, Sustainable Mine Closure, Mining Law, Coal Policy, India & Meghalaya.

I. INTRODUCTION

a. Need

Having lived in India for so many years if there is something that I can talk about for sure is that rental and housing prices keeps growing. The housing prices keeps on increasing at an effective rate on daily basis. After the housing crisis in the year 2008, the housing prices grew at higher rate especially in the various housing brands.

In the year 2016 in the last few months it was wonder to hear that housing prices of Mumbai had fallen the most in the last 4 years. In fact, median resale prices for condos and coops fell 5.6%, marking the first time there was such fall after the year 2016. The fall has been partly credited to political uncertainty domestically and abroad the Political election of 2014.

So to maintain the transparency among customers and also the comparison can be made easy through this model. If any customer is finding the price of the house higher than the price predicted by the model then he/she can simply cross verify the prices or reject the house. This model will ultimately remove dealers and third party sellers from the picture and in turn would increase the productivity for the real estate business.

b. Area of Deployment

Now-a-days, there is full influence of internet over every part of the country. No matter cities or villages, everyone is having a device called mobile with a good speed internet connection. So we thought to integrate our model in to an application or website using Python web app flask in collaboration with HTML and CSS. Every user across the country can login in to the application and provide user credentials and if validated successful then the user can be navigated to home page. In the home page user would be prompted to enter the parameters such as area of the house, parks around house, type of view around house, plants and green surroundings, no of rooms, stories in the house and various other parameters which plays key role in deciding the prices for the house. Once user enters the necessary and required parameters then role of model comes in to the picture and model predicts the prices with the help of regression algorithms and random forest algorithm over which our model has been trained. So, in order...
to access the Pricing app one only needs a device and an internet connection. He can login and can check the prices for the property.

II. SCOPE

To use our Pricing app, there are few steps which user needs to follow

• Registration:

First the user needs to register himself on the app by providing the necessary details.

• Login

Once user has registered on the application then he can login to his account using specific credentials as created by him.

• Using the Pricing application:

After logging in, User can now access the portal and can provide the required parameters for the model to predict the prices using the various machine learning algorithms.

a. ARCHITECTURE OVERVIEW

The project mainly involves two components. The first one is to develop a model using machine learning algorithms and finally train it over a huge data set so that it produces more accurate results. The second part is to integrate this model in an application called pricing app. To develop the model we would be using jupyter notebook as IDE and python as programming language to implement machine learning algorithms. On the other hand to develop application we would be using Django and sql server for database.

b. Stage-1 Data Exploration

Data exploration is the primary step in process of data analysis and consists of all the essential characteristics of a dataset including its size, accuracy, pattern etc. This is mostly performed by data analysts using visual analytics tools but various advanced software can also be used. Python is among one of those which provides various tools for analyzing data. Before conducting analysis over data gathered from various data sources and collected in data warehouses, we must know how many cases are in our dataset, variables, missing values and hypothesis.

c. Stage-2 Data visualization

Data visualization is the graphical view of the information and taking out the meaningful insights from it. By using visual tags like graphs, charts, maps, data visualization tools gives accessible ways to observe trends and outliers in the data. In the world of data driven these tools are very essentials to work upon massive information and data produced every day.
d. **Stage-3 Data selection**

Data selection is the process of extracting the finest data type and source as well as suitable tools to collect data. The method of finding relatable data for a project can affect the integrity of data. The main aim of data selection is the extraction of data type and source. We also need to segregate our data into training and testing data. Training data is that over which our model is trained while testing data is over which our model is tested and it’s accuracy is tested.

<table>
<thead>
<tr>
<th>Price</th>
<th>Area_Sqm</th>
<th>Bedrooms</th>
<th>Latitude</th>
<th>Longitude</th>
<th>PricePerSqM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13300000</td>
<td>74.32</td>
<td>2</td>
<td>19.210200</td>
<td>72.864891</td>
</tr>
<tr>
<td>1</td>
<td>9000000</td>
<td>55.74</td>
<td>1</td>
<td>19.167700</td>
<td>72.949300</td>
</tr>
<tr>
<td>2</td>
<td>9000000</td>
<td>43.68</td>
<td>1</td>
<td>19.130600</td>
<td>72.873816</td>
</tr>
<tr>
<td>3</td>
<td>9000000</td>
<td>78.97</td>
<td>2</td>
<td>19.129005</td>
<td>72.825052</td>
</tr>
<tr>
<td>4</td>
<td>10000000</td>
<td>464.51</td>
<td>5</td>
<td>19.075014</td>
<td>72.907571</td>
</tr>
</tbody>
</table>

**III. SYSTEM ARCHITECTURE**

Let’s think about our data for a moment. Each user will need to create an account for himself and then needs to login using his credentials. After that he needs to provide some input parameters called features based on which our model is going to predict the prices for the house. The input parameters are going to be stored in the backend into the database and finally would be consumed by machine learning model for prediction.

However our main focus will be on the creation of robust and effective machine learning model using regression techniques and random forest algorithm. Finally we will decide which one is best based on the results we get.

a. **Python**

Python is one of the most widely used programming language in scientific and numeric computing. It has large set of libraries which makes our work much more easier in the field of analytics and visualization. Scipy is a library containing package for mathematics and calculation. Pandas is a data analysis and modelling library.
b. Libraries
The libraries used in this project includes Pandas, NumPy, Matplotlib, Seaborn, Scikit Learn etc. These are the libraries supported by python and would be useful in implementing the project. Apart from this we would also be using Jupyter Notebook as an IDE for this project.

c. Models Used
i. Regression Model
Linear Regression is one of the machine learning algorithm working on concept of supervised learning. It basically performs a task of regression. Regression models a target prediction value based on different independent variables. It is mostly used for creating the relationship between forecasting and variables.

![Regression Model](image1)

ii. Random Forest Regression Model
A random forest is again one of the popular machine learning algorithm which performs both regression and classification with help of multiple decision trees and bootstrap aggregation. Bootstrap aggregation is also known as bagging. In the bagging process each decision tree is trained on a different data sample. Here the process of sampling is observed with replacement. The key idea behind this is to aggregate various decision trees in finding the final output rather than relying on individual decision trees.

![Random Forest Regression Model](image2)

d. Dataset
Here we have scrapped the data from the 99acres.com website which is one of the leading real estate websites operating in India. We need to segregate our data into test data and train data. Based on the various input features like crim rate, area, no of rooms, the model is trained and tested.
IV. FINAL RESULT

Based on the performance measure calculated by RMSE (Root Mean Squared Error), we can say that the regression algorithm provides far more better results as compared to random forest algorithm.

Linear Regression displays the best performance for this dataset and can be used for the deployment purpose. We can observe the performance of the model with the following graph.

![Graph showing RMSE for Linear Regression (LR) and Random Forest Regressor (RFR)](image)

a. Model

The model is finally trained using Linear Regression algorithm. The performance measure of the linear regression is far more correlating as compared to Random forest algorithm. Hence, we would be deploying the model based on linear regression over the pricing app for the end users to effectively use it.

b. Template

Python web app flask’s template is a simple file containing text which can produce a text-based format like HTML and XML. The template contains tags and variables. Variables will be replaced by the results when template is evaluated. Tags controls the template’s logic. We also can modify the variables using filters. It means that a lowercase filter would change the variable from uppercase into the lowercase.

V. CONCLUSION

So we can say that our aim is achieved as we successfully ticked all our parameters as stated earlier. Initially we collected the data and then we worked over data by cleaning it and exploring it. After that we performed various analysis over the data to figure out all the meaningful insights from the data. We also tried to visualize the data using various visualizing tools such as maps, graphs, charts etc. After visualizing our data we segregated it into two parts for testing and training purpose. After this, we started to train our model using the two very popularly known algorithms Linear regression and random forest algorithm.
However, afterwards we found that the performance measure of the linear regression was much more significant in predicting the values as compared to random forest algorithm.

For the application we would be using python frontend. In the testing part, we tested our model by giving a wide framework for the user interface and login interface. The predicted prices for the house are reflected to the end user once the input parameters are given to the model.

c. View

![Image of model interface]

The above model is deployed in collaboration with python web app Flask in collaboration with HTML and CSS. The predicted results were very much correlating with the actual results. The RMSE value was observed close to 0.502 for the linear regression model. Hence Linear regression model is considered as best for predicting the prices.

In the final step of deploying this model into an application called price app, we used python web app flas in collaboration with HTML and CSS. To store the user credentials and input features we used database of Sql Server. The user first would be prompted to create an account and after that user can login using his credentials. If the login is successful then the user would be prompted to homepage with an effective UI. The user can provide all the input parameters as asked by the app and then can press enter to check the results. The model provides the correlating price values as compared to actual values.

Hence, we can easily say that the above model would be of great importance for the real estate as well as for the customers who gets easily involved in the trap of brokers.

FUTURE SCOPE

This system is built using a python web app flask in collaboration with HTML and CSS. We are very much aware of the fact that house forecasting is key aspect in the real estate business. Based on the data we came across that population is increasing very rapidly and simultaneously the demand for housing is also increasing proportionally. This entire system would help the real estate as well as customers from getting involved into trap of brokers. No price prediction needs to be done manually and it would save time, cost, human labor for the real estate business. Thus, it would always be in a great demand and this application and model would serve the purpose.

REFERENCES