**AI MINIPROJECT**

**HOUSE PRICE PREDICTION FOR BANGALORE CITY**

**SEM-5**

**Group members: -**

P Poojitha – BT19CSE027

Sneha Madiri – BT19CSE030

D Prathika Yadav – BT19CSE083

G Sampath – BT19CSE120

K Maheswar reddy – BT19CSE123

**Introduction: -**

**House** is the primary need to any human-being. They accumulate all their hard-earned money and put into their life dream of having a house. Before investing their money into any house, they need to investigate about the prices of the houses in their desired area. Our main try and will is to help them know the rough estimation of the house at a particular area.

We use some machine learning models to do this task. In this report, we shall clearly demonstrate the process we followed to get this task done.

**Note: -** This is not any formal research paper or anything, this is just a formal report of what we have done to get this task fulfilled.

**Main steps in predicting the prices of houses given a dataset: -**

1. Procure a good dataset of houses of Bangalore city. (We got this one from Kaggle.com)
2. Perform data cleaning on the dataset u procured.
3. Do feature engineering on the dataset and refine the dataset into model-specific way.
4. Split the dataset into training and testing data.
5. Now provide the training dataset to machine learning model of your wish. See the results using testing data and decide the best model and store that model into some file (of type pickle or joblib).
6. Then start preparing code for the web application. (front-end, back-end, connecting front and back-ends)
7. Finally test whether the website is working fine.

**NOTE: -** Our website is only locally hosted, not global web hosting. Not included going through previous research papers as an internal step of the project (but it was very useful).

**Detailed explanation of each step: -**

For good understanding of below steps I am just printing few rows of our dataset.

**DATA CLEANING: -**

Data cleaning refers to providing data in a way such that it can be provided to the machine learning model directly. Attributes present in the dataset are

“Area\_type, availability, location, size, society, total\_sqft, bath, balcony, price”.

1. Here area\_type, availability, society, balcony are the attributes which actually do not contribute much for the prediction.
2. So, as a part of data cleaning we first drop those 4 columns. Then we check for the null valued datapoints. we get 73 null values for bathrooms (highest of all attributes). As the total file contains more than 13000 houses, 73 is a negligible number. So, we simply drop all the rows with null values.
3. In the size column, the bhk values are given as a string. So, we convert them into integer and store them in bhk named column.
4. Some of the total\_sqft values are given as a range (like 1200-1300 sqft). So, we convert them to average of given range. Now we get integer value of sqft for all the total\_sqft column.
5. Next, we add a new column price\_per\_sqft for later use.
6. Location attribute has categorical values or string values, this needs to be dummied to provide the data to ML model. We have a total of 1298 unique locations, so we will get additional 1297 columns when dummied. This is a huge overhead for ML model to calculate. So, we replace all the locations with less than 10 datapoints in the dataset with “others”. Now, we have 241 unique locations.
7. We dummy the location attribute and remove the location attribute.
8. We can also remove size attribute, as we added bhk column.

Now, we have a dataset purely cleaned and is ready to be provided to any ML model.

Our linear regressor gave a score of 0.4555(approx.), which is very very less.

This tells the importance of feature engineering.

**FEATURE ENGINEERING:**

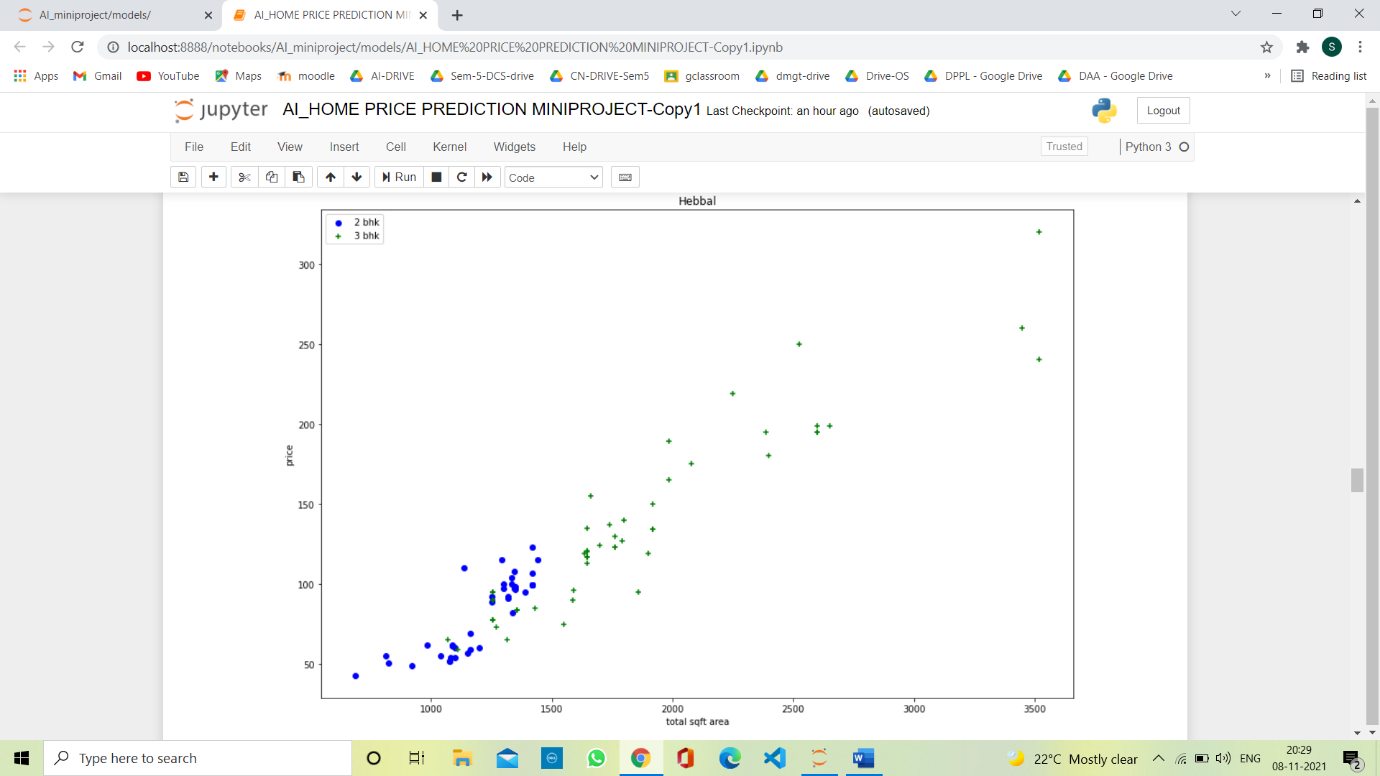
This is a process of removing outliers, datapoints that disobey the business logics.

**Outliers**:

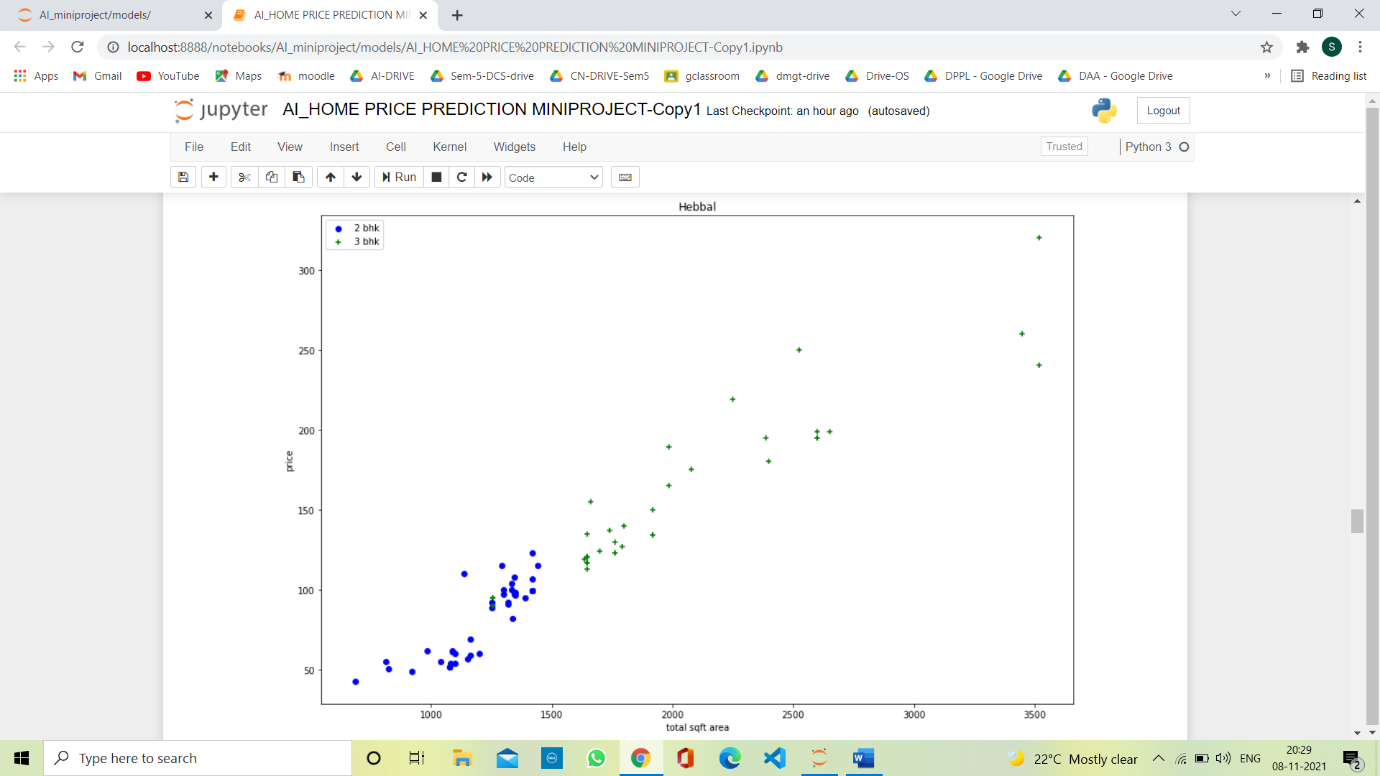
1. We remove the houses which have same total\_sqft but price is greater for the datapoints with less bhk than for more bhk. These are considered as outliers and are to be removed. Note that this is done location wise.

This particular scattered plot is for Hebbal area.

1. Before outlier removal:



1. After outlier removal:



1. All the houses which have price per sqft out of range of (mean+-standard deviation) grouped by location are to be removed. Note that this is done location wise.

**Business logics**:

1. The houses with sqft per bhk less than 300 are considered to be illogical with respect to business perspective. So, these houses will be removed.
2. Houses with no of bathrooms greater than bhk+2 are also considered to be illogical and these will be removed.

Now after all the feature engineering, the no. of rows in dataset are nearly 7000.

Now the dataset is ready to be provided to any ML model.

**TRAIN AND TEST DATASET:**

We have to split the dataset into training and testing dataset with some proportion. Training dataset is used to train the ML models and testing dataset is used to test the ML model for its accuracy in predicting the price.

We used train\_test\_split and Kfold cross validation techniques for splitting the dataset.

**ML MODELS:**

This is the most awaited part of the project. Considering all the research papers we read, we selected to go with linear regression model, lasso regressor, ridge regressor, Decision tree regressor for predicting this continuous variable “price of the house”.

We used GridSearchCV, to find the best parameters for each model and their best scores.

The results:

|  |  |  |
| --- | --- | --- |
| **model** | **Best score** | **Best parameters** |
| Linear regression | 0.849768 | {'normalize': False} |
| Lasso regression | 0.701200 | {'alpha': 1, 'selection': 'cyclic'} |
| Decision Tree Regressor | 0.724357 | {'criterion': 'mse', 'splitter': 'random'} |
| Gradient Boosting | 0.825392 | {'learning\_rate': 0.364, 'n\_estimators': 130} |
| |  | | --- | | **Ridge regression** | | |  |  | | --- | --- | | **0.849818** |  | | **{'alpha': 0.1, 'max\_iter': 25, 'tol': 0.05}** |

Clearly, Linear, Ridge regression won the race with fine margins among all the models.

But out of curiosity our team decided in having a hybrid regressor of different models and check if it gives a higher score than linear regression.

Some of the hybrid regression results:

|  |  |  |
| --- | --- | --- |
| **Model 1 proportion** | **Model 2 proportion** | **Result** |
| Linear regression **0.9** | ridge regression **0.1** | 0.81011 |
| Linear regression **0.05** | ridge regression **0.95** | 0.81036 |
| Linear regression **0.95** | Lasso regression **0.05** | 0.8099 |
| Lasso regression **0.05** | ridge regression **0.95** | 0.81016 |

Clearly, all the proportions are not able to beat individual models. So, we decided to go with individual models. Amongst all the hybrid regressors, 0.95 ridge and 0.05 linear regressions proportion mix gave the best result.

**Server.py:**

Here we just create a flask server and server provides methods for returning location names and predicted price for some given values of independent variables.

**Client.py:**

Client.py has the original implementation of getting location names and getting predicted price which are used by server.py.

**Web Interface:**

Here, we write the HTML, CSS code for “web interface looks” and JS file for reading those inputs from HTML interface and posting them to flask server and getting the result from flask server and displaying it on the interface.

**Conclusion:**

So, hereby we conclude that ridge regressor is the best model for our Bangalore dataset. Finally, we import that model into a pickle file for later use in server files. We also import the locations present in the cleaned dataset for using them in the interface (where user selects his desired location from the drop-down list). Among all the ML models and hybrid models, ridge regression gave better results for this particular dataset.

**COMPARISION WITH OTHER PREVIOUS WORKS:**

|  |  |
| --- | --- |
| **Paper** | **Best score** |
| https://www.kaggle.com/vinodkumarreddyv/banglore-house-price-prediction | 0.845 |
| https://www.irjet.net/archives/V8/i9/IRJET-V8I934.pdf | No score but will address about this during presentation……. |
| **Ours** | 0.849 |