e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

APPLICATION OF ADVANCE REGRESSION TECHNIQUES IN REAL ESTATE PRICE PREDICTION

Dr. B Balamurugan[*1], Shubham Pathak[*2], Ved Pratap Pundhir[*3], Vaibhav Rathore[*4]

[*1]Galgotias University, Associate Professor CSE, Greater Noida UP 201308, India.

[*2,3,4] Galgotias University, CSE, Greater Noida UP 201308, India.

ABSTRACT

When you ask a home buyer to describe their dream home, the basement ceiling height or proximity to an east-west railroad are unlikely to come up first. The evidence from a playground competition, on the other hand, suggests that price agreements are influenced by much more factors than the number of bedrooms or the existence of a white-picket fence. This competition requires you to estimate the final price of each home using 79 explanatory variables that explain (almost) every aspect of Ames, Iowa residential homes. This data science project series will walk you through the process of building a real estate price prediction website from start to finish. We'll start by creating a model with sklearn and linear regression using the home prices dataset from kaggle.com. The next step is to build a flask server in Python that serves http requests using the saved model. The third part is an html, CSS, and javascript-based website that allows users to input details such as square footage, bedrooms, and so on, and then calls a Python Flask server to get an approximate price. Almost all data science concepts will be addressed during this course, including data loading and cleaning, outlier detection and removal, feature engineering, dimensionality reduction, grid search cv for hyper parameter tuning, k fold cross validation, and so on.

Keywords: k cross fold Validation , grid search.

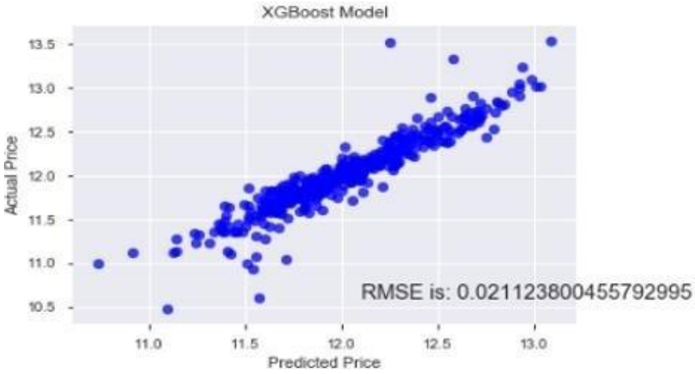
I. INTRODUCTION

We need a proper forecast on real estate and houses in the housing market. We can see a mechanism that runs in the properties purchasing and selling. For most people, buying a house would be a lifetime ambition. There are a lot of people making big mistakes right now when it comes to purchasing properties. Most people are buying properties they haven't seen from people they don't know through ads, and one of the most popular mistakes is buying properties that are too costly but aren't worth it.

Real estate may be required to include a valuation of the property. Many different players in the commercial core, such as property agents, appraisers, assessors, mortarboard lenders, traders, developers, and gurus, perform a quantitative measure of profit[5]. Reserve owners, lenders, and others are also included. That will be used to determine the value of a company. The word requisition comes from the Latin word requisition, which From alleging valuation schemes, as well as methods that represent the value of the property and the circumstances under which it is provided for. Property could well be on the way to being exchanged in the open market under a variety of conditions and circumstances; however, many people are unaware of the current situation and begin to lose money. There are numerous factors that influence how real estate prices are determined. Every aspect of this report will be examined in order to forecast the actual price of a house.

II. ACQUISITION OF DATA

Before getting into in depths of any project , the very first thing we need is the dataset to work upon . Therefore the first step that constitutes any project is acquisition of data . In this case, we must determine which houses to consider and how we will forecast their prices. It's a crucial part of the project that must be handled with care. Obviously, we cannot use a large dataset because it would be impractical, so for this project, we have chosen 20 houses to estimate prices.

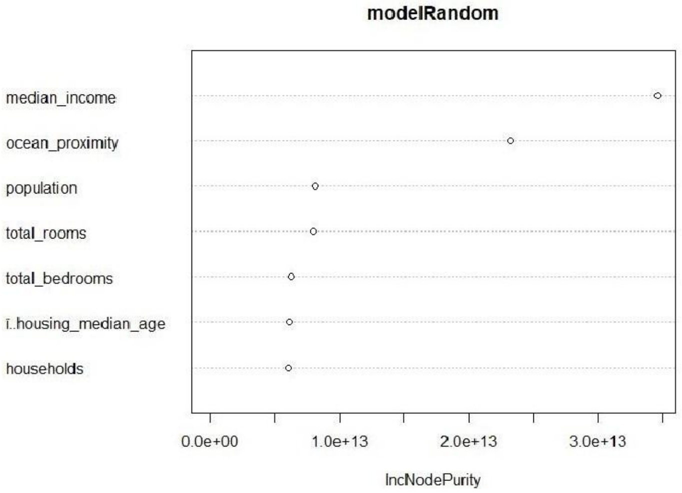


The standard deviation of the errors that have crept in when predicting the values is known as the Root Mean Square Error (RMSE). Residues are used to calculate the distance between the points and the regression line.. The RMSE calculates the distribution of these points, or simply the spread of these points. In other words, RMSE indicates how densely or clustered the points are along the best-fit axis. The root mean square error is widely used in many fields to validate the outcomes of different experiments.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)^2}{n}}$$

III. DATA CLEANING AND INTEGRATION

The first version of data cleaning focuses on locating and correcting inaccurate data. Since we are using various prediction techniques, the cleaning process differs, but the ultimate goal is to achieve greater precision. Before loading the data into the machine learning models, the data from the repository should be corrected for errors and null values. This will ensure that the prediction accuracy is large. It's possible that some real estate data details are missing. It's likely that only a few details are given, and that some of the information provided is inaccurate.. As a consequence, it's critical to double-check all information and keep just what's necessary. As a result, data cleaning is critical because it eliminates a lot of redundant data, which only adds to the task's difficulty. To minimise inconsistency, null values are also omitted.



1 Feasibility Analysis

It refers to how practical it is to develop a system that defines or rather estimates the price of the real Estates in the current market . Given the fact that a major population is paying either very less or large sums of money in order to attain a particular property , it is very necessary that a particular system needs to be followed .

In terms of technical feasibility , we are going to use the services that are available readily and for free on internet . Therefore it is quite feasible to follow this approach and use all the available tools .

2 Tools

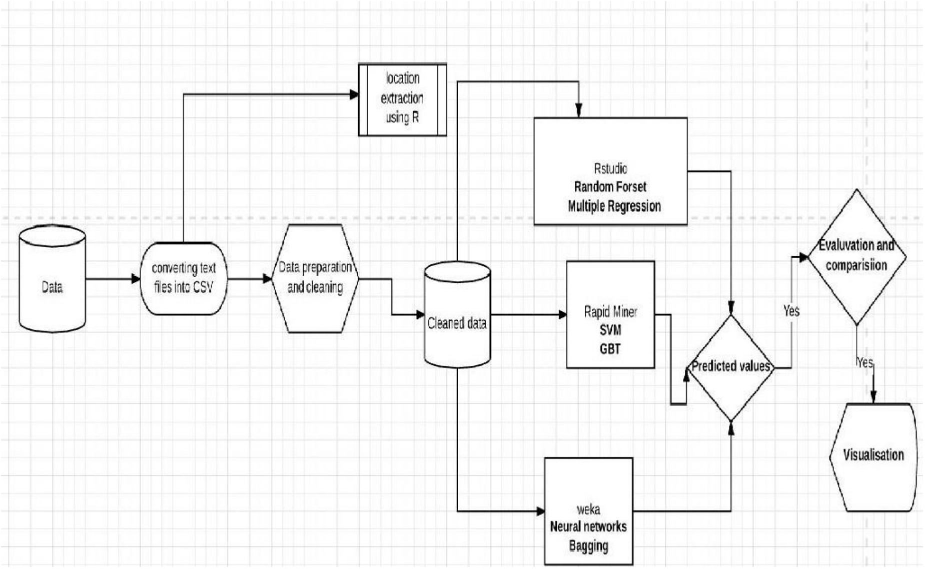
Technology and tools wise this project covers

- Python which is a programming language.
- Data cleaning with Numpy and Pandas
- Data visualisation with Matplotlib
- Learn how to create models
- As an IDE, use Jupyter notebook, Visual Studio Code, and PyCharm.
- HTML/CSS/Javascript for UI, Python flask for http server

3 Data Preprocessing

This procedure involves methods for removing any null or infinite values that could influence the system's accuracy. Formatting, washing, and sampling are the key steps[4]. The cleaning process is used to remove or correct any missing data, as well as data that is incomplete.

Sampling is a method in which sufficient data are used to minimise the algorithm's running time. The preprocessing is carried out in Python.

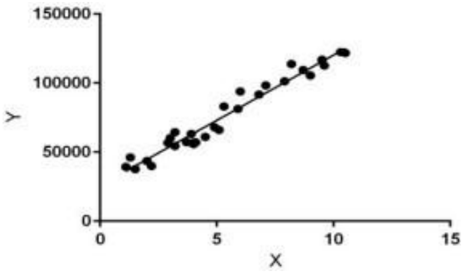


4 Regression Algorithm

Linear regression is a supervised learning machine learning algorithm. It carries out a regression mission. [three]

Based on independent variables, regression models a target prediction value. It is primarily used in forecasting and determining the relationship between variables. Different regression models vary in terms of the type of relationship they consider between dependent and independent variables and the number of independent variables they employ. In simple terms, regression is a machine learning technique that aids in the prediction of future events by gathering information from current data and about the relationships between our target parameter and a number of other parameters. According to this concept, the cost of a home is determined by a variety of factors such as the number of bedrooms, bathrooms, location, and society. We will know the actual value of a property in a given geographical area by adding all of these parameters to machine learning. The simple idea of regression is observing the relation between output and input and then applying the resultant function to our data , Choosing an Algorithm - There are various methods of doing Regression Analysis. The sole purpose of everything is to predict the outcome with accuracy . A most commonly used method is by calculating the r factor which is nothing but difference between an actual and a predicted value raised to power two.

There can be a situation in which our model may overfit if we use same data for learning and checking the accuracy . Putting it it more simpler words , the model will work fine for a given data set but will fail entirely whenever a new data set is observed . A suggested solution to this kind of problem is dividing the dataset into two parts and using one part for testing the accuracy and other for testing . By using this technique we can know if our model overfits .



Linear regression is used to estimate the value of a dependent variable (y) based on the value of an independent variable (x). As a result of this regression technique, a linear relationship between x (input) and y (output) is discovered (output). As a result, the term Linear Regression was coined.

In the diagram above, X (input) represents work experience and Y (output) represents a person's salary. For our model, the regression line is the best fit line.

We are given the following model to train:

x: input training data

y: the data's labels (supervised learning)

It matches the best line to predict the value of y for a given value of x while training the model. By determining the best 1 and 2 values, the model obtains the best regression fit rows.

1st: interception 2: the x-coefficient

$$y = \theta_1 + \theta_2.x$$

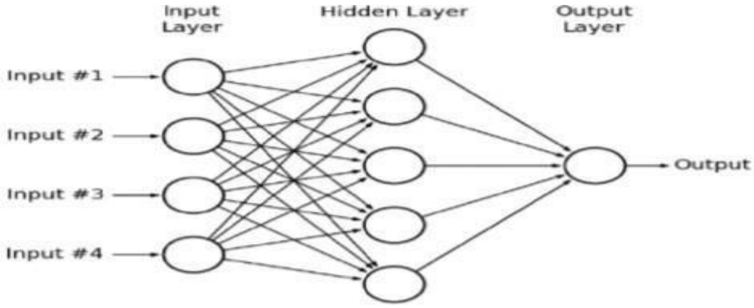
We get the best match line after we find the best 1 and 2 values. So, when we use our model to predict the value of y for the input value of x, it will predict the value of y.

How can the 1 and 2 values be modified to get the best fit line? (J) Cost Function:

The model aims to forecast y value in such a way that the error gap between expected and true value is as small as possible by achieving the best-fit regression line. As a result, it is important to change the 1 and 2 values in order to find the best value that minimises the difference between the expected y value (pred) and the real y value (y) .

5 Neural Networks

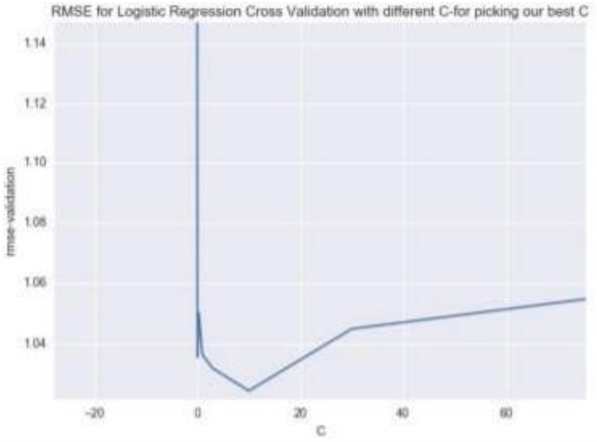
MLP stands for multilayer perceptron in its complete form. It has the same structure as a single layer perceptron with one or more hidden layers. In these, the input layer is directly connected to the hidden layer, and the inputted values are also present in the perceptron. The perceptrons can now be measured against a specific set of values, and if the observed values match the inputted values, the model's output is deemed adequate, and no additional innovations or system changes are required [1]..



Cross validation is an algorithm of machine learning technology that is used to evaluate models on a given data set The process starts with spliting the data into k equal folds and that is why this process is often termed as k fold cross validation

The value of k can be variable and suppose k =20 then the model is said to be 20 fold cross validation. This process is generally used to how a model performs on an unseen data . the performance of the model is very essential . The system is divided into k folds and each part is tested against an input and the result obtained is tested across remaining k-1 folds . It mainly tests the performance of the system

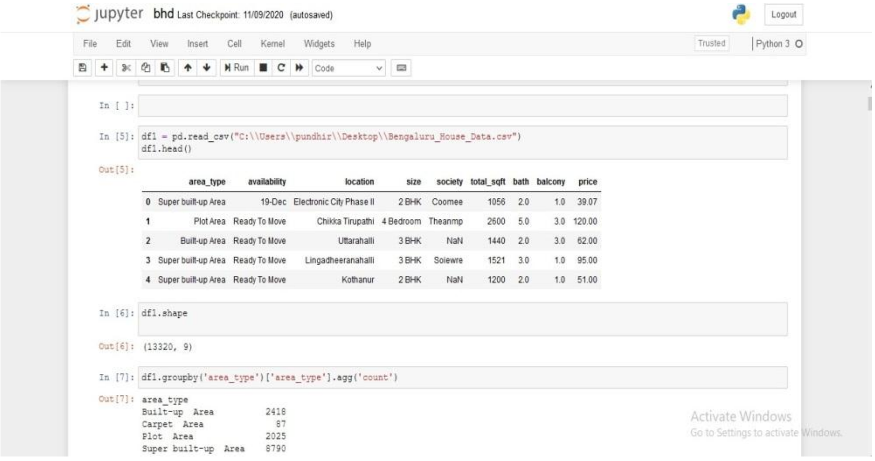
.. It is simple to use , popular and easy to understand method as it generally results in lesser biased and optimistic model estimate as compared to other algorithms.[2]



6 Our System

Working in data science necessitates the use of a platform that allows us to undertake the most efficient operations in order to achieve the best possible results. For data science, Jupyter Notebook was the first choice. Taking a set of sample sets is important because we used a functional platform. We're working on predicting real estate prices. Because importing data in a proper format is necessary, we utilise the 'pandas' package. Because our sample data contains thirteen thousand entries, the following operations are necessary.

a . Data cleansing and data frame creation



```
In [ ]:

In [5]: df1 = pd.read_csv("C:\\Users\\pundhir\\Desktop\\Bengaluru_House_Data.csv")
        df1.head()
```

Out[5]:

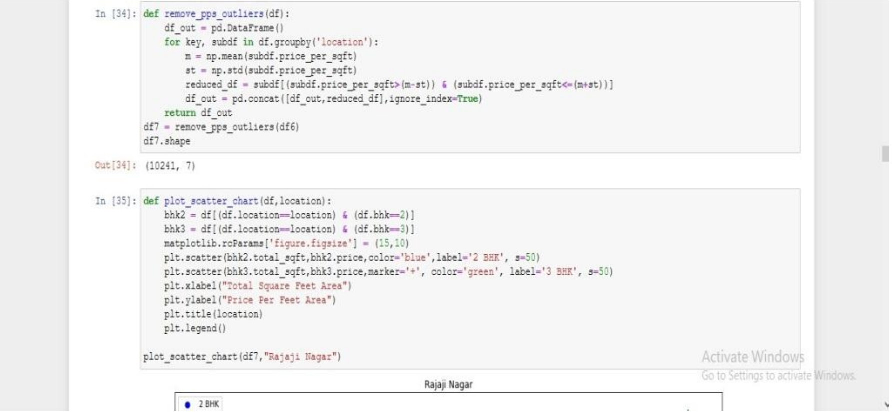
	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

```
In [6]: df1.shape

Out[6]: (13320, 9)

In [7]: df1.groupby('area_type')['area_type'].agg('count')

Out[7]: area_type
        Built-up  Area          2418
        Carpet  Area              87
        Plot  Area              2025
        Super built-up  Area    8790
```

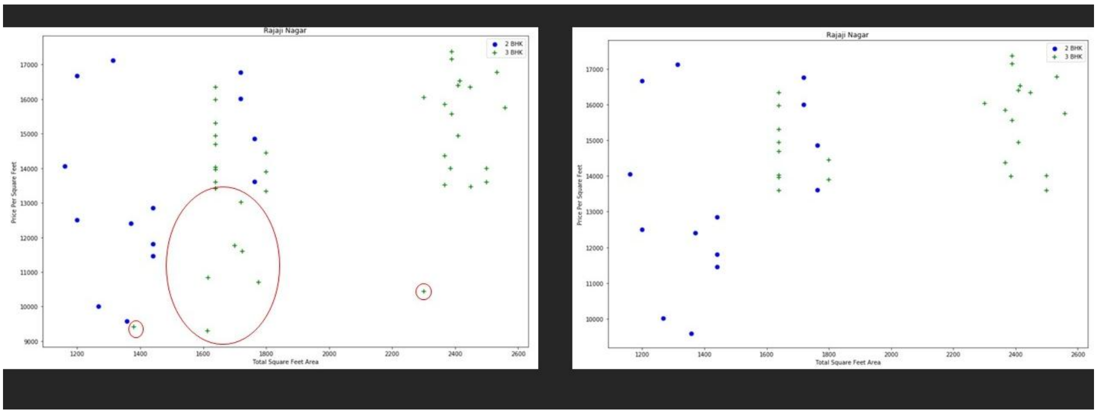
A DataFrame is a two-dimensional data structure in which data is organised in rows and columns in a tabular format. Pandas DataFrames make it simple to manipulate your data. You may pick and choose columns and rows, as well as alter your data. Using the, we can delete a row or a column.

b . Removing outliers

When you talk to your company manager (who is an expert in real estate) as a data scientist, he would tell you that a typical square ft per bedroom is 300 (i.e. a 2 bhk apartment is at least 600 sqft). If you have a 400 sqft flat with two bedrooms, for example, it is odd and can be eliminated as an anomaly. We'll keep our minimal threshold per bhk at 300 sqft to eliminate such outliers.



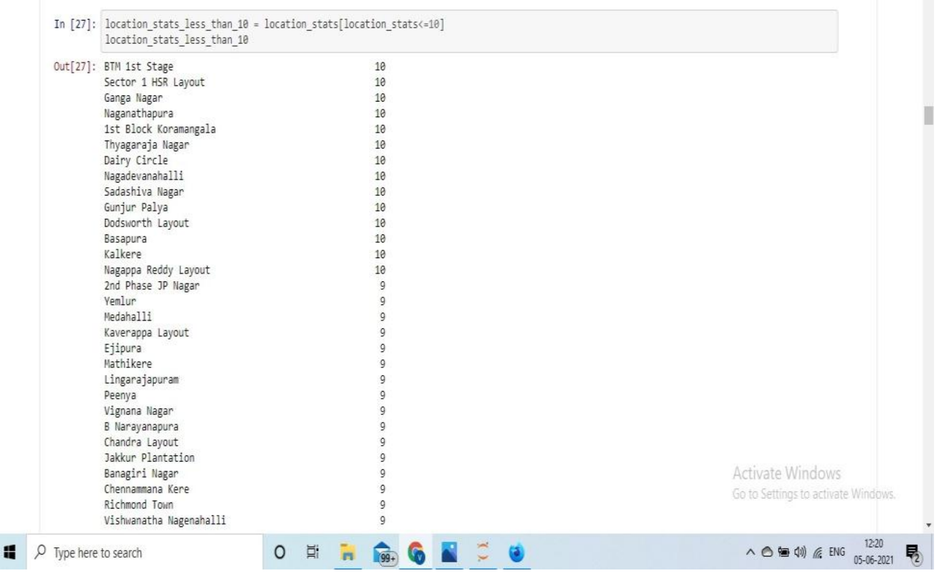
```
In [34]: def remove_pps_outliers(df):
             df_out = pd.DataFrame()
             for key, subdf in df.groupby('location'):
                 m = np.mean(subdf.price_per_sqft)
                 st = np.std(subdf.price_per_sqft)
                 reduced_df = subdf[(subdf.price_per_sqft>(m-st)) & (subdf.price_per_sqft<=(m+st))]
                 df_out = pd.concat([df_out,reduced_df],ignore_index=True)
             return df_out
         df7 = remove_pps_outliers(df6)
         df7.shape

Out[34]: (10241, 7)

In [35]: def plot_scatter_chart(df,location):
             bhk2 = df[(df.location==location) & (df.bhk==2)]
             bhk3 = df[(df.location==location) & (df.bhk==3)]
             matplotlib.rcParams['figure.figsize'] = (15,10)
             plt.scatter(bhk2.total_sqft,bhk2.price,color='blue',label='2 BHK', s=50)
             plt.scatter(bhk3.total_sqft,bhk3.price,marker='+', color='green', label='3 BHK', s=50)
             plt.xlabel("Total Square Feet Area")
             plt.ylabel("Price Per Feet Area")
             plt.title(location)
             plt.legend()

         plot_scatter_chart(df7,"Rajaji Nagar")
```



Before and after removing Outliers can be seen in the image above .

c . Dimensionality Reduction

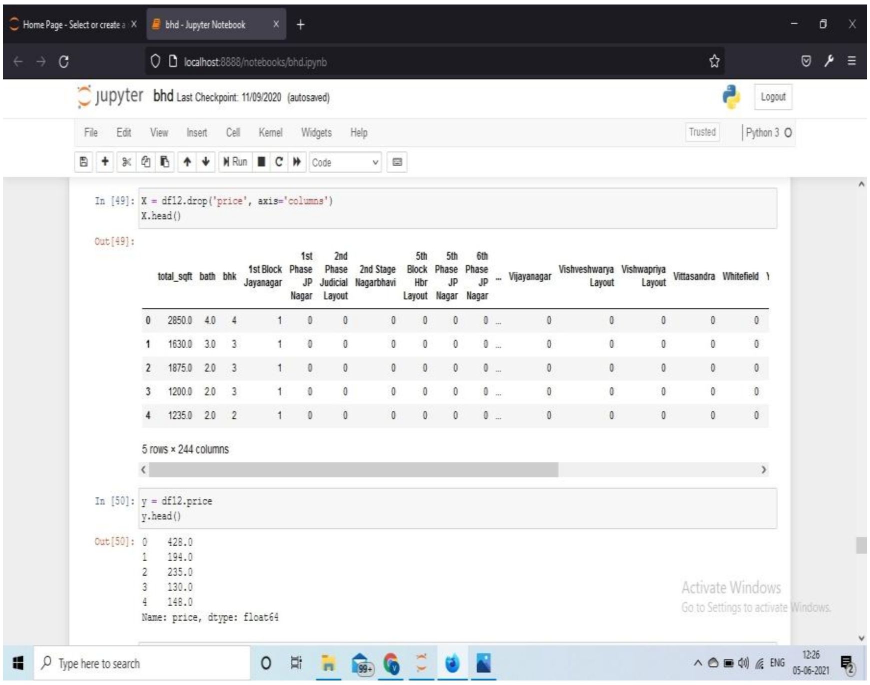
Any location with fewer than ten data points should be labelled as "other." The number of categories can be drastically decreased in this manner. It will help us have less dummy columns later on when we execute one hot encoding.



```
In [27]: location_stats_less_than_10 = location_stats[location_stats<=10]
         location_stats_less_than_10

Out[27]: BTM 1st Stage                  10
         Sector 1 HSR Layout            10
         Ganga Nagar                    10
         Naganathapura                  10
         1st Block Koramangala          10
         Thyagaraja Nagar               10
         Dairy Circle                   10
         Nagadevanahalli                10
         Sadashiva Nagar                10
         Gunjur Palya                   10
         Dodsworth Layout               10
         Basapura                       10
         Kalkere                        10
         Nagappa Reddy Layout           10
         2nd Phase JP Nagar              9
         Yemlur                          9
         Medahalli                       9
         Kaverappa Layout                9
         Ejipura                         9
         Mathikere                       9
         Lingarajapuram                  9
         Peenya                          9
         Vignana Nagar                   9
         B Narayanapura                  9
         Chandra Layout                  9
         Jakkur Plantation               9
         Banagiri Nagar                  9
         Chennammana Kere                9
         Richmond Town                   9
         Vishwanatha Nagenahalli         9
```

d . Model building

1 We will use K fold cross validation to measure accuracy of our Linear Regression model .



```
In [49]: X = df12.drop('price', axis='columns')
         X.head()
```

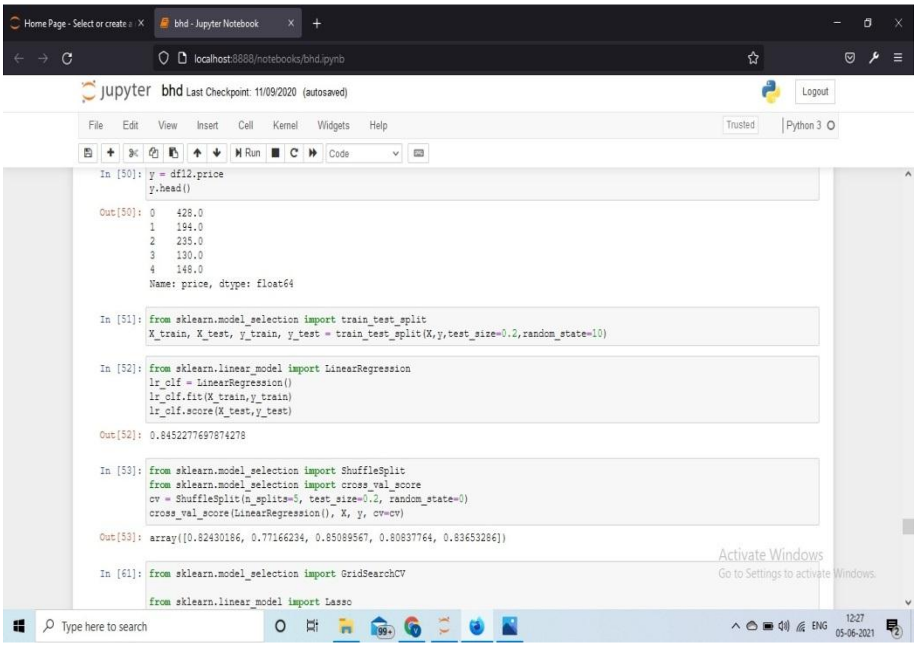
Out[49]:

	total_sqft	bath	bhk	1st Block Jayanagar	1st Phase JP Nagar	2nd Phase Judicial Layout	2nd Stage Nagarbhavi	5th Block Hbr Layout	5th Phase JP Nagar	6th Phase JP Nagar	...	Vijayanagar	Vishveshwarya Layout	Vishwapriya Layout	Vittasandra	Whitefield
0	2850.0	4.0	4	1	0	0	0	0	0	0	...	0	0	0	0	0
1	1630.0	3.0	3	1	0	0	0	0	0	0	...	0	0	0	0	0
2	1875.0	2.0	3	1	0	0	0	0	0	0	...	0	0	0	0	0
3	1200.0	2.0	3	1	0	0	0	0	0	0	...	0	0	0	0	0
4	1235.0	2.0	2	1	0	0	0	0	0	0	...	0	0	0	0	0

5 rows × 244 columns

```
In [50]: y = df12.price
         y.head()

Out[50]: 0    428.0
         1    194.0
         2    235.0
         3    130.0
         4    148.0
         Name: price, dtype: float64
```

IV. RESULTS AND DISCUSSION

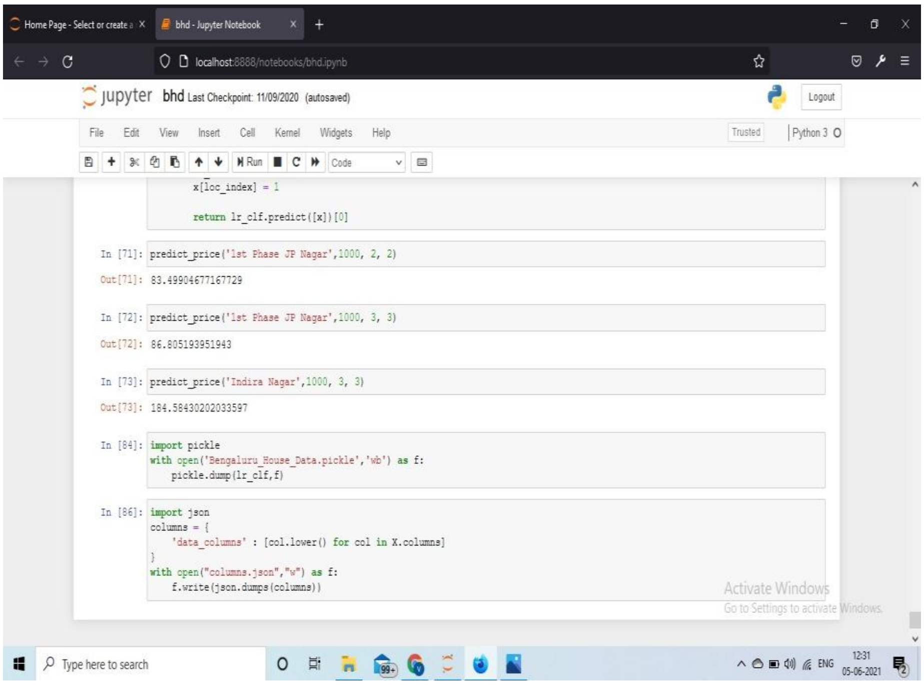


```
In [50]: y = df12.price
         y.head()

Out[50]: 0    428.0
         1    194.0
         2    235.0
         3    130.0
         4    148.0
         Name: price, dtype: float64

In [51]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,random_state=10)

In [52]: from sklearn.linear_model import LinearRegression
         lr_clf = LinearRegression()
         lr_clf.fit(X_train,y_train)
         lr_clf.score(X_test,y_test)

Out[52]: 0.8452277697874278

In [53]: from sklearn.model_selection import ShuffleSplit
         from sklearn.model_selection import cross_val_score
         cv = ShuffleSplit(n_splits=5, test_size=0.2, random_state=0)
         cross_val_score(LinearRegression(), X, y, cv=cv)

Out[53]: array([0.82430186, 0.77166234, 0.85089567, 0.80837764, 0.83653286])

In [61]: from sklearn.model_selection import GridSearchCV

         from sklearn.linear_model import Lasso
```

We can observe that we always achieve a score above 80% after 5 iterations. This is a solid start, but we'd like to try a couple different regression techniques to see if we can improve our score even further. Grid Search CV will be used for this.

We may conclude that Linear Regression provides the best score based on the above results. As a result, we'll make advantage of it.

a . Importing as model as pickle file

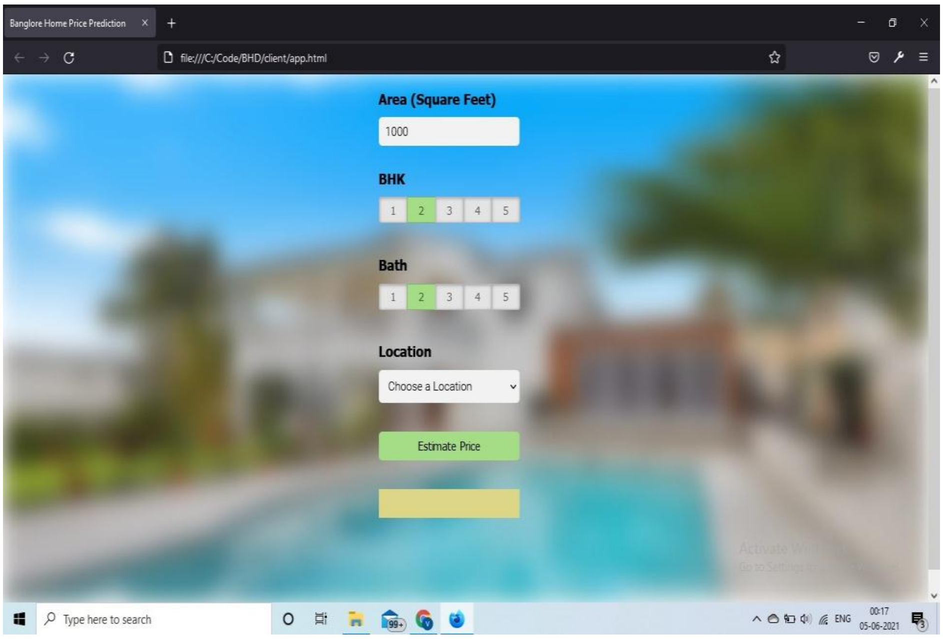


```
        x[loc_index] = 1

        return lr_clf.predict([x])[0]

In [71]: predict_price('1st Phase JP Nagar',1000, 2, 2)

Out[71]: 83.49904677167729

In [72]: predict_price('1st Phase JP Nagar',1000, 3, 3)

Out[72]: 86.805193951943

In [73]: predict_price('Indira Nagar',1000, 3, 3)

Out[73]: 184.58430202033597

In [84]: import pickle
         with open('Bengaluru_House_Data.pickle','wb') as f:
             pickle.dump(lr_clf,f)

In [86]: import json
         columns = {
             'data_columns' : [col.lower() for col in X.columns]
         }
         with open("columns.json","w") as f:
             f.write(json.dumps(columns))
```

bhd - Jupyter Notebook
localhost:8888/notebooks/bhd.ipynb
jupyter bhd Last Checkpoint: 11/09/2020 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

b . Building an UI

Technology we are using are HTML, CSS, JavaScript



Area (Square Feet)

1000

BHK

1 [✓ 2] 3 4 5

Bath

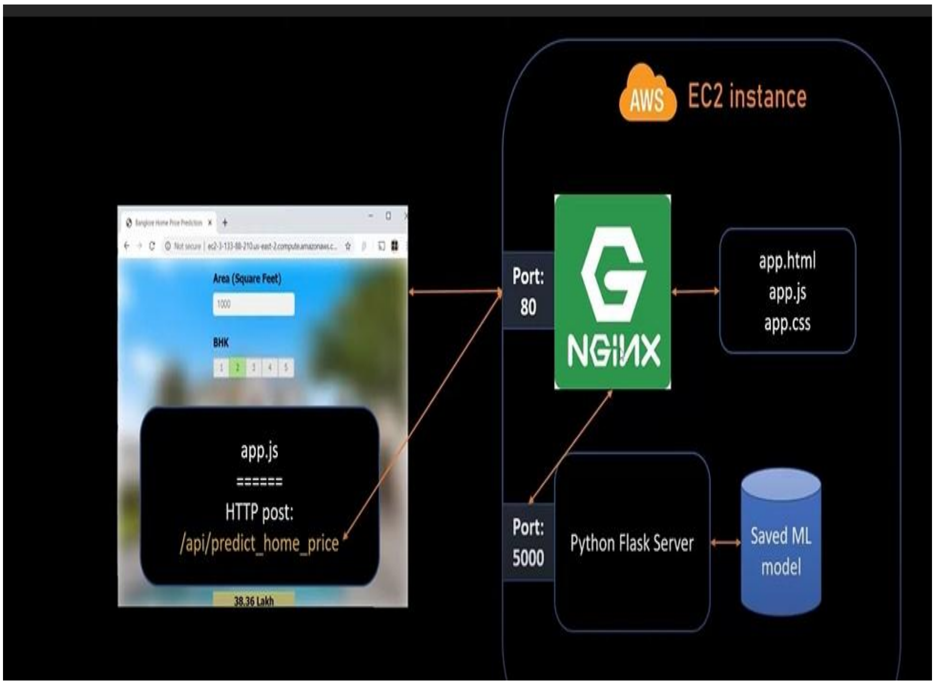
1 [✓ 2] 3 4 5

Location

Choose a Location

Banglore Home Price Prediction
file:///C:/Code/BHD/client/app.html
Estimate Price

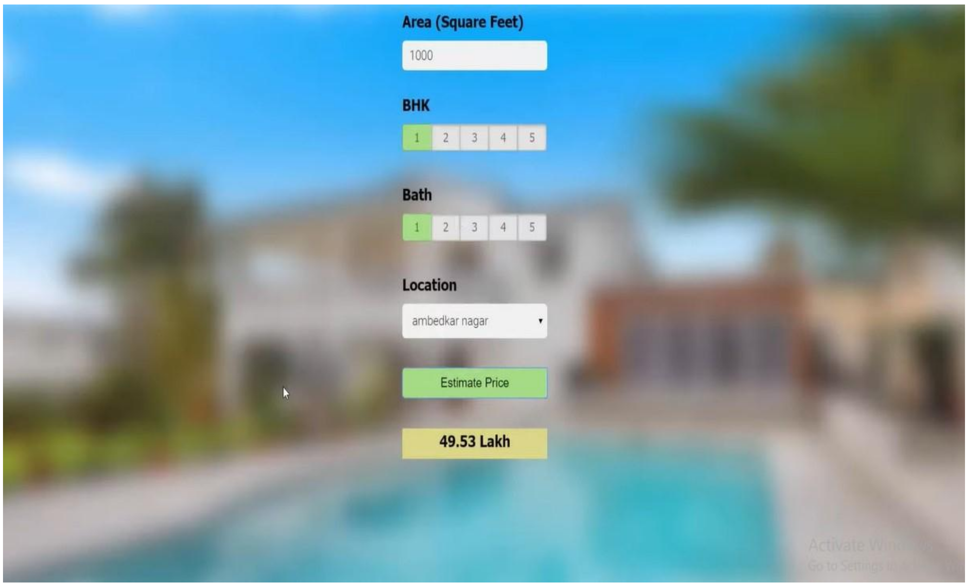
c . Deploy machine learning model to production



Nginx, a reverse proxy, load balancer, mail proxy, and HTTP cache, is a web server that may also be used as a reverse proxy, load balancer, and mail proxy. Igor Sysoev built the software, which was publicly released in 2004. Nginx is a free and open-source web server that is distributed under the conditions of the BSD 2-clause licence. NGINX is used by a significant number of web servers, often as a load balancer.

A reserve call to the NginX server is made when a user clicks the button to forecast the home price from the js file. The architecture employs a reserve proxy system based on NginX to route all requests (excluding API queries) to a Python Flask server operating on the same EC2 instance, which will utilise the saved MLmodel to estimate the price.

d . Functional Website images



Area (Square Feet)
1000

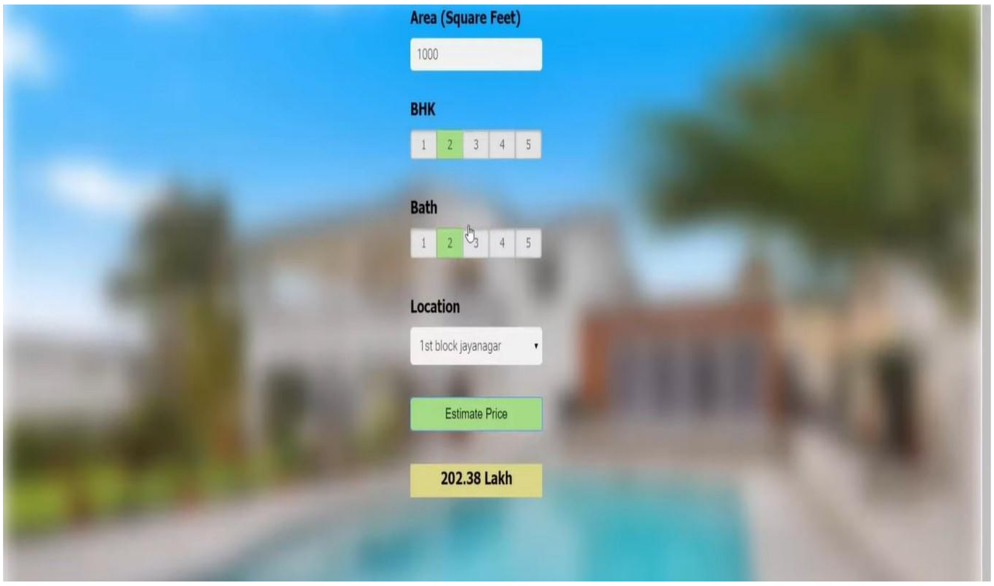
BHK
1 2 3 4 5

Bath
1 2 3 4 5

Location
ambedkar nagar

Estimate Price

49.53 Lakh



Area (Square Feet): 1000

BHK: 1 [✓ 2] 3 4 5

Bath: 1 [✓ 2] 3 4 5

Location: 1st block jayanagar

202.38 Lakh

V. CONCLUSION

We want to develop a hassle-free experience platform for those with no prior experience in construction or home building, planning, and reducing the work efforts of professionals who are already familiar with the real estate environment. Individuals would be able to dive deep into real estate expertise, engage with experts, and get an estimate on their plan.

VI. REFERENCES

[1] Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.

[2] Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

[3] https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression 14c4e325882a

[4] https://analyticsindiamag.com/5-ways-handle-missing-values-machine- learning-datasets/

[5] https://www.investopedia.com/articles/mortgages-real- estate/11/valuing-real-estate.asp