REAL ESTATE PRICE PREDICTION SYSTEM

Internship Member (s)

Sl. No.	Reg. No.	Student Name
1	19ETCS002093	RADHA BAI G N
2	19ETCS002313	KAVYA
3	19ETCS002075	NAMRATHA M

Supervisors:1. Prof.Pallavi R Kumar

Nov/Dec - 2022

B. Tech. in Computer Science and Engineering

FACULTY OF ENGINEERING AND TECHNOLOGY

M. S. RAMAIAH UNIVERSITY OF APPLIED SCIENCES

Bengaluru -560 054

FACULTY OF ENGINEERING AND TECHNOLOGY

Certificate

RADHA BAI GN

This is to certify that the Internship project titled "REAL ESTATE PRICE PREDICTION" is a bonafide work carried out in the Department of Computer Science and Engineering by MS. RADHA BAI G N Bearing Reg. No. 19ETCS002093 in partial fulfilment of requirements of the Course curriculum of 7thSem Computer Science and Engineering of Ramaiah University of Applied Sciences.

Nov/Dec - 2022

Name of Mentor: prof. Pallavi R Kumar
Designation: professor
Place: Bangalore
Date:

Certificate

KAVYA

This is to certify that the Internship project titled "REAL ESTATE PRICE PREDICTION" is a bonafide work carried out in the Department of Computer Science and Engineering by MS. KAVYA bearing Reg. No. 19ETCS002313 in partial fulfilment of requirements of the Course curriculum of 7thSem Computer Science and Engineering of Ramaiah University of Applied Sciences.

Nov/Dec - 2022

Name of Mentor: prof. Pallavi R Kumar
Designation: professor
Place: Bangalore
Date:

Certificate

NAMRATHA M

This is to certify that the Internship project titled "REAL ESTATE PRICE PREDICTION" is a bonafide work carried out in the Department of Computer Science and Engineering by MS. NAMRATHA M Bearing Reg. No. 19ETCS002313 in partial fulfilment of requirements of the Course curriculum of 7thSem Computer Science and Engineering of Ramaiah University of Applied Sciences.

Nov/Dec - 2022

Name of Mentor: prof. Pallavi R Kumar
Designation: professor
Place: Bangalore
Date:

Acknowledgements

I would like to thank my supervisor **prof. Pallavi R Kumar** for the patient guidance, encouragement and advice he has provided throughout my time as his student. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly.

I am truly thankful to my Head of Department
Dr. PUSHPHAVATHI T P who gave me this opportunity to do the project on this topic (Exploratory Data Analysis for Time Series Datasets) as well as my sincere thank you

To **Dr.Dilip Kumar Mahanty** Dean of the Faculty, for the continuous encouragement.

I would like to thank Prof. Gangadhar and prof. Deepak Vardhan for guiding us to successfully complete the intern project.

Abstract

House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets.

Machine learning techniques are applied to analyse historical property transactions in India to discover useful models for house buyers and sellers.

Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of Mumbai.

Moreover, experiments demonstrate that the Multiple Linear Regression that is based on mean squared error measurement is a competitive approach.

We propose to implement a house price prediction model of Bangalore, India. It's a Machine Learning model which integrates Data Science and Web Development. We have deployed the app on the Heroku Cloud Application Platform. Housing prices fluctuate on a daily basis and are sometimes exaggerated rather than based on worth. The major focus of this project is on predicting home prices using genuine factors. Here, we intend to base an evaluation on every basic criterion that is taken into account when establishing the pricing. The goal of this project is to learn Python and get experience in Data Analytics, Machine Learning, and Al.

Summary

For every project the literature review will give clear idea and it will serve as the base line here most of the authors have concluded that artificial neural networks have the more influence in predicting but in the real world the other algorithms should be also taken into consideration.

By conducting this study, it helps to know about both the pros and cons and it had helped me to successfully implement the project.

It is tough in today's real estate world to store large amount of data and extract them according to one's need. The data extracted should be useful. This system uses linear regression algorithm and makes the model look optimal. The data is used in the most efficient way using the required algorithm. The model helps to fulfil users need by maintaining the accuracy of estate choice and reducing the risk of investment.

In addition to this, adding more databases of other cities will help the customer to explore more estates and reach a more accurate decision. Along with this, factors which affect the house price like recession shall be added. In-depth details of the properties should be added to provide more information to the user.

Also, using larger number of data sets will increase the accuracy of the model more. Different model can also be used so that the calculation time decreases and whole process can be carried out in ease.

Table of Contents

List of Tables

List of Figures

Nomenclature

We have not used any nomenclature

Abbreviation and Acronyms

AI – Artificial Intelligence

ML – Machine Learning

MA – Moving Average

AR – Auto Regressive

HPI – House Price Index

1. Introduction

In this project, Machine Learning (ML) is a vital aspect of present-day business and research. It progressively improves the performance of computer systems by using algorithms and neural network models. Machine Learning algorithms automatically build a mathematical model using sample data also referred to as training data which form decisions without being specifically programmed to make those decisions.

People and real estate agencies buy or sell houses, people buy to live in or as an investment and the agencies buy to run a business. Either way, we believe everyone should get exactly what they pay for. over-valuation/under-valuation in housing markets has always been an issue and there is a lack of proper detection measures. Broad measures, like house/Real-estate price-to-rent ratios, give a primary pass. However, to decide about this issue an in-depth analysis and judgment are necessary. Here's where machine learning comes in, by training an ML model with hundreds and thousands of data a solution can be developed which will be powerful enough to predict prices accurately and can cater to everyone's needs.

The primary aim of this paper is to use these Machine Learning Techniques and curate them into ML models which can then serve the users. The main objective of a Buyer is to search for their dream house which has all the amenities they need. Furthermore, they look for these houses/Real estates with a price in mind and there is no guarantee that they will get the product for a deserving price and not overpriced. Similarly, a seller looks for a certain number that they can put on the estate as a price tag and this cannot be just a wild guess, lots of research needs to be put to conclude a valuation of a house.

Additionally, there exists a possibility of under-pricing the product. If the price is predicted for these users, this might help them get estates for their deserving prices not more not less.

Literature Survey

Real estate has become more than a necessity in this 21st century, it represents something much more nowadays. Not only for people looking into buying Real Estate but also the companies that sell these Estates. 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 real estate price can affect various household investors, bankers, policymakers, and many. Investment in the real estate sector seems to be an attractive choice for investments. Thus, predicting the real estate value is an important economic index.

Most of the literature study is based on articles with full text online, open access articles and peer-reviewed publications from database search engine Summon, and the search websites; the Research Gate publications and the Towards Data Science collection instead of textbooks and chapters of books. The literature study endeavours to construct a robust basis on regression techniques, regularisation, and artificial neural network in machine learning and on how it can precisely be applied to house prices prediction.

There is a vast amount of work that is focused on training models to detect patterns in datasets to predict what the future output could be. However, there are researches where the authors use different machine learning algorithms with a combination of pre-processing data methods.

A research was conducted in 2017 by Lu, Li and Yang [20]. They examined the creative feature engineering and proposed a hybrid Lasso and Gradient boosting regression model that promises better prediction. They used Lasso in feature selection. They used the same dataset as the one used in this study. They did many iterations of feature engineering to find the optimal number of features that will improve the prediction performance. The more features they added, the better the score evaluation they receive from the website Kaggle. Hence, they added 400 features on top of the 79 given features. Furthermore, they used Lasso for feature selection to remove the unused features and found that 230 features provide the best score by running a test on Ridge, Lasso and Gradient boosting.

In 2016, Jose Manuel Pereira, Mario Basto and Amelia Ferreira da Silva performed a study to examine three methods [21]. Lasso, Ridge and Stepwise

Regression implemented in SPSS to develop an empirical modelfor predicting corporate bankruptcy. They defined two types of errors. The first error is the percentage of failed enterprises predicted well by the model. The second error is the percentage of good enterprises predicted failed by the model. The results of this study showed that the lasso and ridge algorithms tend to favour the category of the dependent variable that appears with heavier weight in the training set when they are compared to the stepwise algorithm implemented in SPSS.

A study was accomplished in 2017 by Suna Akkol, Ash Akilli, Ibrahim Cemal [22], where they did a comparison of Artificial neural network and multiple linear regression for prediction. In their study, the impact of different morphological measures on live weight has been modelled by artificial neural networks and multiple linear regression analyses. They used three different back-propagation techniques for ANN, namely Levenberg-Marquardt, Bayesian regularisation, and Scaled conjugate. They showed that ANN is more successful than multiple linear regression in the prediction they performed.

The literature study gives an overview of the articles that are related to this study, the feature engineering methods that have been used in this study. As well as evaluation metrics that are used to measure the performance of the algorithms. In addition, the factors that have been used in the local dataset.

CONCLUSION AND FUTURE SCOPE

Buying your own house is what every human wish for. Usingthis proposed model, we want people to buy houses and real estate at their rightful prices and want to ensure that they don't get tricked by sketchy agents who just are after their money. Additionally, this model will also help Big companies by giving accurate predictions for them to set the pricing and save them from a lot of hassle and save a lot of precious time and money. Correct real estate prices are the essence of the market and we want to ensure that by using this model.

The system is apt enough in training itself and in predicting the prices from the raw data provided to it. After going through several research papers and numerous blogs and articles, a set of algorithms were selected which were suitable in applying on both the datasets of the model. After multiple testing and training sessions, it was determined that the XGBoost Algorithm showed the best result amongst the rest of the algorithms. The system was potent enough for Predicting the prices of different houses with various features and was able to handle large sums of data. The system is quite user-friendly and time-saving.

The supplementary feature that can be added to our proposed system is to avail users of a full-fledged user interface so there can be multiple functionalities for users to use with the ML model for numerous locations. Also, an Amazon EC2 connection will take the system even further and increase the ease of use. Lastly, developing a well-integrated web application that can predict prices whenever users want it to will complete the project.

Background Theory

Multiple Linear Regression

Multiple Linear Regression (MLR is a supervised technique used to estimate the relationship between a dependent variable and more than one independent variables. Identifying the correlation and its cause-effect helps to make predictions by using these relations).
To estimate these relationships, the prediction accuracy of the model is essential; the complexity of the model is of more interest. However, Multiple Linear Regression is prone to many problems such as multi-collinearity, notes, and overfitting, which effect on the prediction accuracy.

Regularised regression plays a significant part in Multiple Linear Regression because it helps to reduce variance at the cost of introducing some bias, avoid the overfitting problem and solve ordinary least squares problems.
There are two types of regularisation techniques L1 norm (in absolute deviations) and 12 norms (least squares). L1 and L2 have different cost functions regarding model complexity

ARIMA

The ARIMA model predicts a given time series based on its own past values. It can be used for any non-seasonal series of numbers that exhibits patterns and is not a series of random events. For example, sales data from a clothing store would be a time series because it was collected over a period of time.

For data scientists, the ARIMA model is a vital tool for providing accurate forecasts across a wide range of disciplines. For example, a manufacturing company uses an ARIMA model to drive business planning, procurement and production goals. Errors in the forecast could cause significant disruption in the supply chain and production activities of the company. Accurate predictions can help lower costs and meet customer expectations with greater efficiency.

The ARIMA model can also be used for climate studies, such as predicting greenhouse gas concentrations or monitoring the weather patterns.

Aim and Objectives

2.1 Title

Real Estate Price Prediction

2.2 Aim

The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted.

2.3 Objectives

The objective wasto forecast the price of a specific apartment based on market pricing while accounting for various “features” that would be established in the following sections

This application will help the buyer invest in estate without approaching the agent. This also decreases the risk in investment. The current trend of buying is hectic and bit expensive as the customer has to roam to various places and also need to pay commission to the agent or the manager. Hence, we will design a website using the data science techniques to overcome all the drawbacks of the currently used system. We are implementing the following in our website:

i) Location/area-based search

ii) Approx. cost of property considering the different attributes and factors.

2.4 Scope of the Project

Bangalore is a dense city with a lot of population the demand for land and property is gone sky rocketing in recent years this proposed model serves the purpose of finding the essential values and prices of the properties

In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected from houses. This study utilizes machine learning algorithms as a research method. House price prediction can help determine the selling price of a house and can help to arrange the right time to purchase a house. There are various factors that influence the price of a house which include physical conditions, concept and location.

2.5 Need and Motivation

Having lived in India for so many years if there is one thing that I had been taking for granted, it's that housing and rental prices continue to rise. Since the housing crisis of 2008, housing prices have recovered remarkably well, especially in major housing markets.

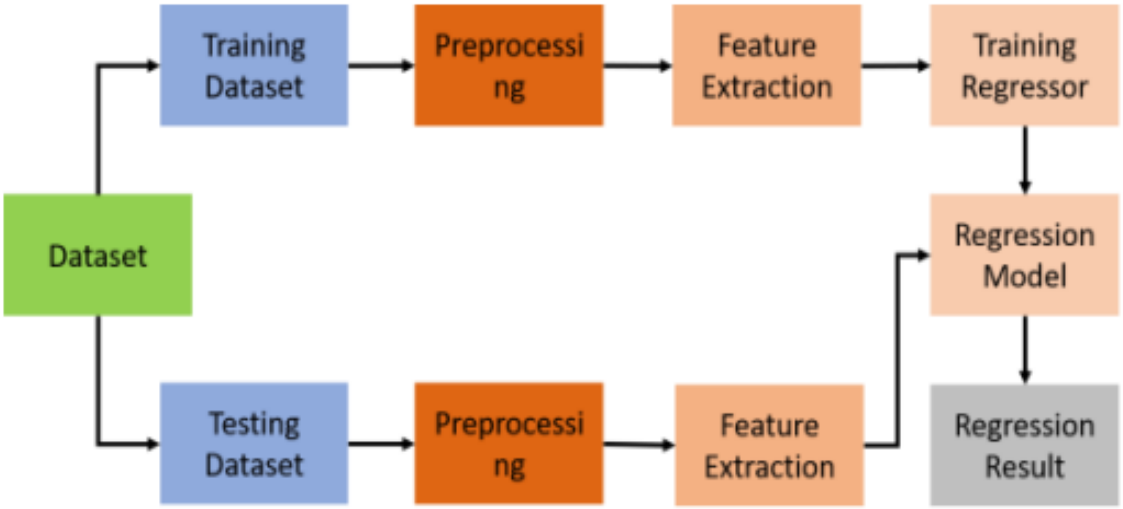
However, in the 4th quarter of 2016, I was surprised to read that Bombay housing prices had fallen the most in the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first time there was a decline since Q1 of 2017.

The decline has been partly attributed to political uncertainty domestically and abroad and the 2014 election. So, to maintain the transparency among customers and also the comparison can be made easy through this model.

If customer finds the price of house at some given website higher than the price predicted by the model, so he can reject that house.

2.a

Methods and Methodology/Approach to attain each objective:



The above block diagram is the traditional Machine Learning Approach. It consists of two sections: the training and the testing. The training has the following components: the label, input, feature extractor, and the machine learning algorithm. The testing section has the following components in it: the input, feature extractor, the regression model, and the output label. Input: The input consists of data collected from various sources. Feature Extractor: Only important features which affect the prediction results are kept. Other unnecessary attributes are discarded, like ID or name. Features: After feature extraction only, some inputs are considered which largely contribute to the prediction of the model. Machine Learning Algorithm: The ML Algorithm is the method by which an AI system performs its task, and is most commonly used to predict output values from given input values. Regression is one of the main processes of machine learning. The Regression Model: The regression model consists of a set of machine-learning methods that allow us to predict a label variable (y) based on the values of one or more attribute/feature variables (x). Briefly, the goal of a regression model is to build a mathematical equation that defines y as a function of the x variables.

Discussion and Results

Introduction to the topic:

Machine Learning (ML) is a vital aspect of present-day business and research. It progressively improves the performance of computer systems by using algorithms and neural network models. Machine Learning algorithms automatically build a mathematical model using sample data also referred to as training data which form decisions without being specifically programmed to make those decisions.

People and real estate agencies buy or sell houses, people buy to live in or as an investment and the agencies buy to run a business. Either way, we believe everyone should get exactly what they pay for. over-valuation/under-valuation in housing markets has always been an issue and there is a lack of proper detection measures. Broad measures, like house/Real-estate price-to-rent ratios, give a primary pass. However, to decide about this issue an in-depth analysis and judgment are necessary. Here's where machine learning comes in, by training an ML model with hundreds and thousands of data a solution can be developed which will be powerful enough to predict prices accurately and can cater to everyone's needs.

The primary aim of this paper is to use these Machine Learning Techniques and curate them into ML models which can then serve the users. The main objective of a Buyer is to search for their dream house which has all the amenities they need. Furthermore, they look for these houses/Real estates with a price in mind and there is no guarantee that they will get the product for a deserving price and not overpriced. Similarly, A seller looks for a certain number that they can put on the estate as a price tag and this cannot be just a wild guess, lots of research needs to be put to conclude a valuation of a house.

Additionally, there exists a possibility of under-pricing the product. If the price is predicted for these users, this might help them get estates for their deserving prices not more not less.

3. Label:

The label is the output obtained from the model after training. The data obtained from the dataset is given as a training input first and the relevant training features are extracted. These training features are preprocessed to get a normalized dataset and labeling of the data row is done.

The result from the training dataset is fed to the machine learning algorithm.

The result from the Machine Learning Algorithm is fed to the Regression model, thus producing a trained model or trained regressor. This trained regressor can take the new data that is the extracted feature from the test as input and predict its output label.

4. Design:

Software Requirements:

Coding Language:

Coding software:

- Python3
- HTML
- Python Flask
- JavaScript

Anaconda

Spyder

Subline text 3

4.1. Functional Requirements:

User Interface: The user interface will be a website. The user has to enter all the attributes correctly and in the required format. Proper Forecasting: The system has to properly predict the price of the house according to the input given by the user. Database: Dataset should contain large number of entities so that it will increase the accuracy of the predicted price and suggest a better property.

Proper Forecasting: The system has to properly predict the price of the house according to the input given by the user.

Database: Dataset should contain large number of entities so that it will increase the accuracy of the predicted price and suggest a better property.

4.2 Non Functional Requirements:

Platform Independent: The application would be platform independent if all the requirements are installed in the device.

Performance: The application should have better accuracy and should provide the information in less time.

Capacity: The capacity of the storage should be high so that large amount of data can be stored in order to train the model.

5. SYSTEM DESIGN:

5.1 User interface: The user interface for our project is Website. For this software, the users are the businessman, investors and other people searching for property. They have to enter details about the property they want and then the software will give the accurate predicted value. User can also forecast the predicted value by entering date. In this application, the user has to enter information on website about the user's location such as number of floors, area in sq. feet, location, bhk, furnishing, date for forecasting and budget.

5.2 Data Preparation: To prepare the dataset for the prediction system, some changes were made:

Binary categorical variables are represented using one binary digit (i.e. (Furnishing) 0 = Not Furnished, 1 = Furnished).

Also by using label encoder names of places is to be converted into values as linear regression model is to be trained by using values.

As the price of properties are often quoted in lakhs, we have rounded our dependent variable to nearest thousand which also helps with numerical stability of model.

5.3 DATA SETS:

Dataset is Extracted from kaggle.com by using concept

of Web Scraping for house price prediction purpose

Dataset used for prediction contains names of all the areas in and nearby Bengaluru with their BHK, Sq.ft, Society, availability, area type, bathrooms, balconies and Prices.

It contains 13330 entries which contains over 100 areas and places in Bengaluru.

A	B	C	D	E	F	G	H	I
area_type	availability	location	size	society	total_sqft	bath	balcony	price
Super built	19-Dec	Electronic	2 BHK	Coomee	1056	2	1	39.07
Plot Area	Ready To M	Chikka Tiru	4 Bedroom	Theanmp	2600	5	3	120
Built-up A	Ready To M	Uttarahalli	3 BHK		1440	2	3	62
Super built	Ready To M	Lingadhee	3 BHK	Soiewre	1521	3	1	95
Super built	Ready To M	Kothanur	2 BHK		1200	2	1	51
Super built	Ready To M	Whitefield	2 BHK	DuenaTa	1170	2	1	38
Super built	18-May	Old Airpor	4 BHK	Jaades	2732	4		204
Super built	Ready To M	Rajaji Nag	4 BHK	Brway G	3300	4		600
Super built	Ready To M	Marathaha	3 BHK		1310	3	1	63.25

The two datasets in this paper where various existing machine learning algorithms are applied to the datasets for predicting prices.

The first dataset is from the UCI Machine Learning Repository which concerns housing values in the suburbs of Boston. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. As this paper uses machine learning for price prediction, attribute variables are used to predict the label/price.

5.4.1. Linear Regression:

- In this Project, we have used Linear Regression Algorithm for predicting the current house price.
- The Linear Regression Algorithm accepts two variables Independent variable (X) and Dependent variable (Y).
- We have used sklearn Library for importing Linear Regression model.
- The dataset containing different cities with their features and prices is used for training Linear Regression Model.
- The dataset entities will be divided into two parts 80% for training and 20% for testing.
- Linear Regression model will be trained using X_train Independent variable entries and Y_train Dependent variable entries.
- The trained model will be tested upon the 20% test dataset entities. After training and testing the model will be used for prediction purpose.
- The accuracy for trained linear regression model is 86.67%.

Formula:

$Yi=\beta0+\beta1Xi1+\beta2Xi2+...+\beta pxi+\epsilon$

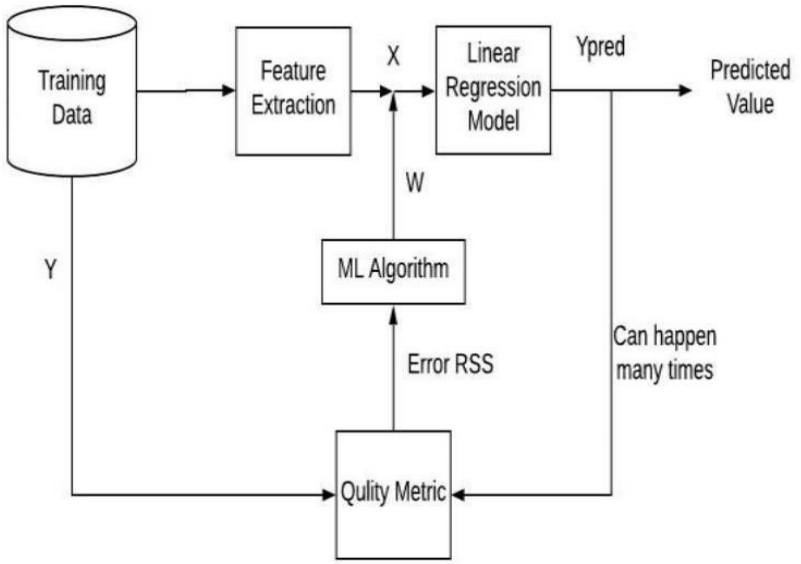
Yi=dependant variables

Xi=independent variables

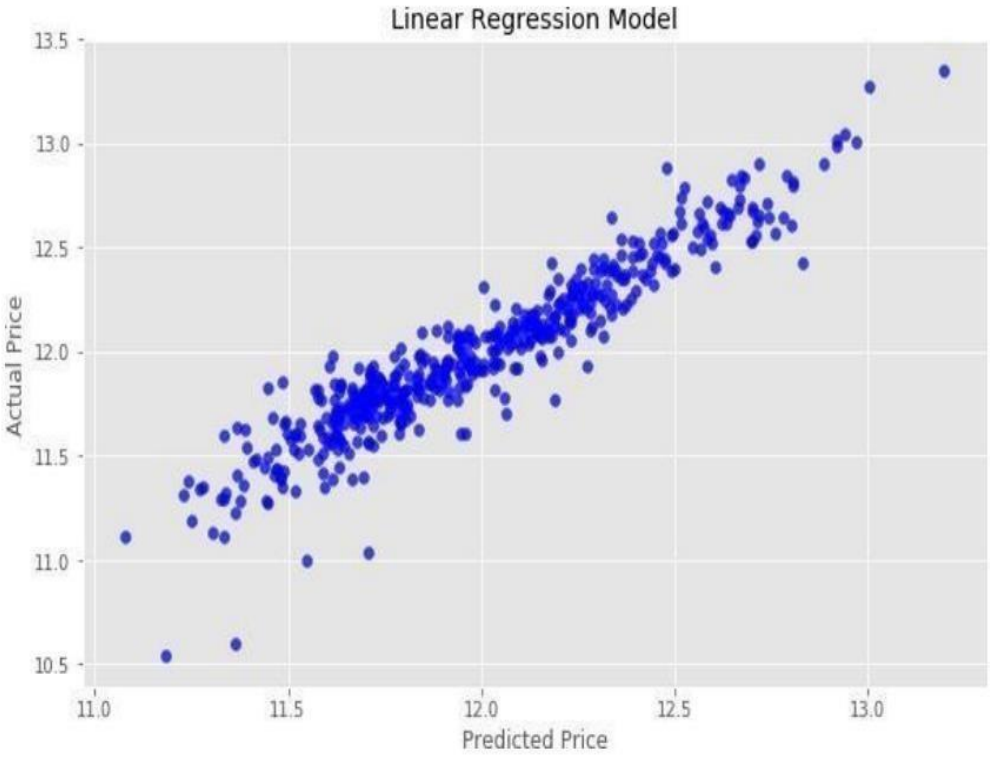
β0= y-intercept (constant term)

ϵ=the model's error

Dependent Variables	Independent Variables
LOCATION(STRING)	PRICE
AREA IN SQ.FT(INT)	
BHK	
BATHROOMS	



PREDICTION MODEL SCATTER



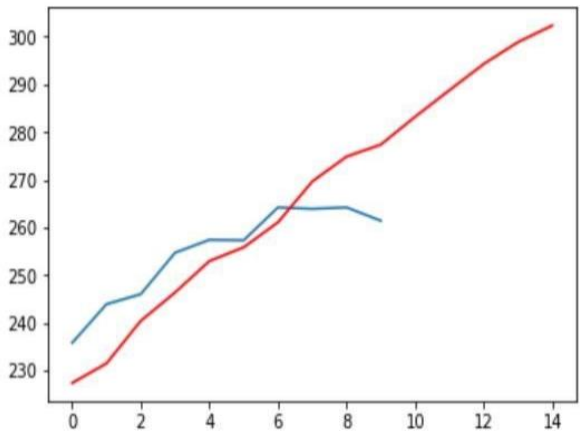
PLOT FOR LINEAR REGRESSION MODE

5.4.2. ARIMA (Auto Regressive Integrated Moving Average Model):

- ARIMA Model is widely used for Forecasting purpose like stock, temperature forecasting, sales predictions etc.
- In this project, the ARIMA is used to forecast house price for a particular the data which is given by the user.
- ARIMA Model is the combination for three methods for forecasting which are Autoregressive (AR) Model, Integrated differencing and Moving Average (MA) Model.

1. Autoregressive(AR) Model: Yt depends only on past values Yt-1, Yt-2, so on. Yt = F(Yt-1,Yt2,Yt3,…) if no.of past values (p) increases then the accuracy of the model increases.
2. Moving Average (MA) Model :Yt depends only on past error terms. Yt=F(Et, Et-1, Et-2,…) the no.of past error terms taken is mostly 0, 1 or 2. The No.of error terms is denoted by ‘q’.

The dataset entities will be divided into two parts 80% for training and 20% for testing. The ARIMA model is imported from stats model library which takes training dataset and order of (p, d, q) as input. After training the Model will be used for forecasting purpose. This project contains ARIMA model with 87% of accuracy.



ACCURACY OF ARIMA

5.5. Model procedure

The trained linear regression model is given the user entered property details as input and model will return predicted value which is pass from flask to website. For Arima model the input will the user entered date. The ARIMA model will give House Price Index (HPI) as output which is converted to House price by using formula

Current House Price Value * Future HPI = Future House Price Value * Current HPI

Then the Forecasted house price is displayed on web by flask.

We will make machine learning model which will use k-fold cross validation and grid search cv to come up with best algorithm and best parameters. Along with this we are using pandas' dummies method to treat the dummy values. While building our model we divide the data set into training and testing data to evaluate the model performance. To find the best optimal model we will use two methods namely k-fold cross validation and Grid-Search-CV which will tell us which is the best algorithm for our model. We have imported the linear regression, lasso and decision tree algorithm. Out of these the best is selected to build the model.

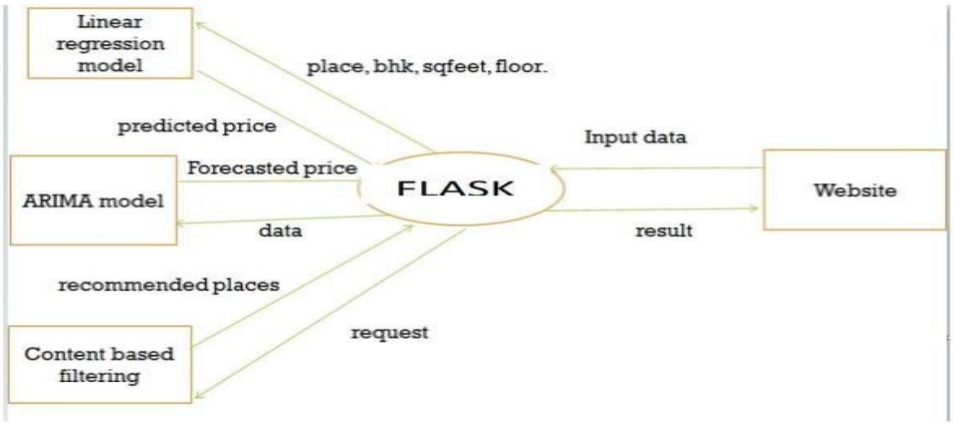
5.6. Connectivity

- The website is connected to backend by using framework called python flask.
- The flask provides a local IP address through which the websites is connected.
- When user enter details about property on website, the IP address provided by flask is used to pass data to flask.

In python flask program, the trained linear regression model is imported by using library joblib and property details fetched from URL is given to the trained model. The output is given as predicted price which is displayed on screen.

- Similarly, User entered date is also fetch from URL for forecasting. This date is given to the imported ARIMA model which gives forecasted HPI(House Price Index) as output.

The forecasted HPI is used to calculate forecasted price which is return to the website.



WORKING OF THE SYSTEM

6. Status of the Work

Content	status
1. Title and Aim	COMPLETED
2. Objectives	COMPLETED
3. Methods and methodology/Block diagram	COMPLETED
4. Design 4.1 Functional requirements 4.2 Non functional requirements	COMPLETED
5. System design 5.1 User interface 5.2 Data preparation 5.3 Datasets 5.4 Methodology 5.4.1 Linear Regression 5.4.2 ARIMA 5.5 Model procedure 5.6 connectivity	COMPLETED
6. Status of work	COMPLETED
7. Results 7.1 Backend outputs 7.1.1 Scatter plot 7.1.2 Bar graphs 7.2 Frontend snapshots	COMPLETED

Content	Status
8. Expected outcomes 8.1 Demonstration of working model 8.2 A new technique 8.3 A patent	COMPLETED
9. Cost estimation	COMPLETED
10. Gantt chart	COMPLETED
11. References	COMPLETED

7.Implementation

```
import pandas as pd import numpy as
np from matplotlib import pyplot
as plt
%matplotlib inline   import matplotlib matplotlib.rcParams["figure.figsize"]=(20,10)


df1=pd.read_csv("D:\ML mini project\Bengaluru_House_Data.csv") df1.head()
df2=df1.drop(['area_type','society','balcony','availability'],axis='columns')
df2.head() df3['bhk']=df3['size'].apply(lambda x:int(x.split(' ')[0]))


def convert_sqft(x):
tokens=x.split('-')        if len(tokens)==2:
            return(float(tokens[0])+float(tokens[1]))/2
        try:
            return float(x)        except:

            return None  df4=df3.copy()
df4['total_sqft']=df4['total_sqft'].apply(convert_sqft)
df4.head(45)
#df4.loc[412]
df5=df4.copy() df5['price_per_sqft']=df5['price']*100000/df5['total_sqft']
df5.head(15)
```


```
df5=df4.copy() df5['price_per_sqft']=df5['price']*100000/df5['total_sqft'] df5.head(15)

df5.location=df5.location.apply(lambda x : 'other' if x in location_stats_less_than_10 else x  )

len(df5.location.unique())


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     def remove_bhk_outliers(df):

exclude_indices=np.array([])     for location

,location_df in df.groupby('location'):

bhk_stats={}         for bhk,bhk_df in

location_df.groupby('bhk'):

bhk_stats[bhk]={

            'mean':np.mean(bhk_df.price_per_sqft),

            'std':np.std(bhk_df.price_per_sqft),

            'count':bhk_df.shape[0]

        }

    for bhk,bhk_df in location_df.groupby('bhk'):

        stats=bhk_stats.get(bhk-1)           if

stats and stats['count']>5:
```

```
exclude_indices=np.append(exclude_indices,bhk_df[bhk_df.price_per_sqft<(stats['mean'])].index.val
ues)    return df.drop(exclude_indices,axis='index')          df8=remove_bhk_outliers(df7) df8.shape

mport matplotlib matplotlib.rcParams['figure.figsize']=(20,10)
plt.hist(df8.price_per_sqft,rwidth=0.8)
plt.xlabel('price_per_sqft') plt.ylabel('Count')

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)

from sklearn.linear_model import LinearRegression
lr_clf=LinearRegression() lr_clf.fit(X_train,y_train)
lr_clf.score(X_test,y_test)

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)
from sklearn.model_selection import GridSearchCV from
sklearn.linear_model import Lasso  from sklearn.tree
import DecisionTreeRegressor

def find_best_model_using_gridsearchcv(X,y):
    algos={
```

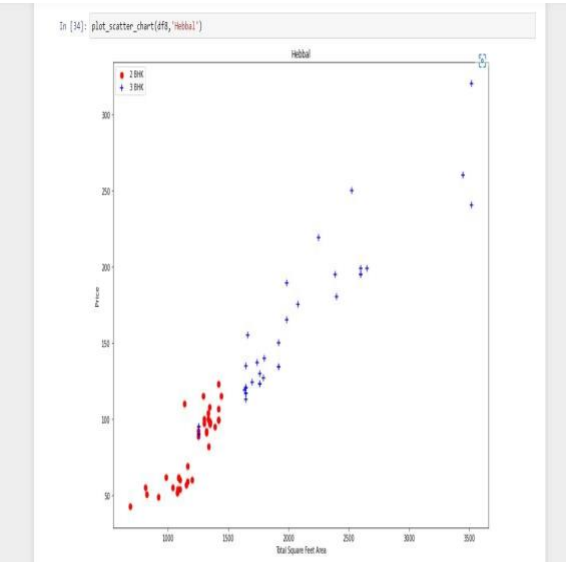

```
    'linear_regression': {
    'model':LinearRegression(),
      'params':{
        'normalize':[True,False]
      }
    },
    'decision_tree':{
      'model':DecisionTreeRegressor(),
      'params':{
        'criterion':['mse','friedman_mse'],
        'splitter':['best','random']
      }
    }

  }   scores=[]
cv=ShuffleSplit(n_splits=5,test_size=0.2,random_state=0)     for
algo_name,config in algos.items():
    gs=GridSearchCV(config['model'],config['params'],cv=cv,return_train_score=False)
    gs.fit(X,y)
scores.append({
      'model':algo_name,
      'best_score':gs.best_score_,
      'best_params':gs.best_params_
    })
```

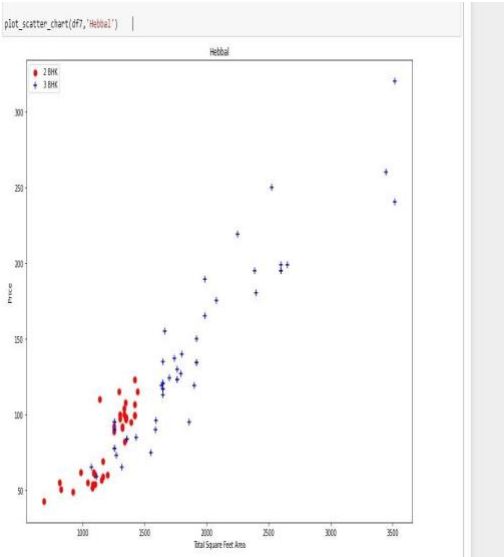
```
    return                                  pd.DataFrame(scores,columns=['model','best_score','best_params'])

find_best_model_using_gridsearchcv(X,y) def predict_price(location,sqft,bath,bhk):

    loc_index =  np.where(X.columns==location)[0][0]


    x = np.zeros(len(X.columns))

x[0] = sqft     x[1] = bath

x[2] = bhk      if loc_index >= 0:

x[loc_index] = 1      return

lr_clf.predict([x])[0]


predict_price('Nagarbhavi',600,1,2)
```

Results

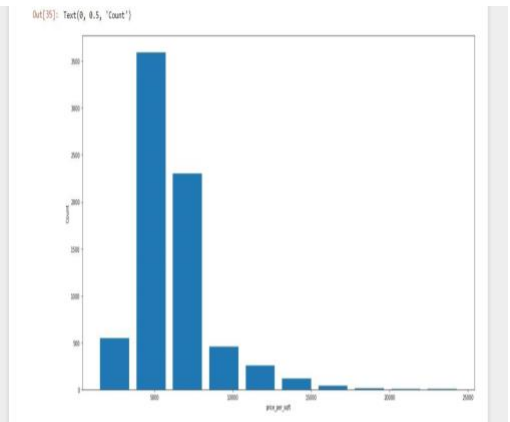
7.1 Backend Outputs: The python code which is executed in Jupiter notebook will give following charts as the outputs for specific conditions.

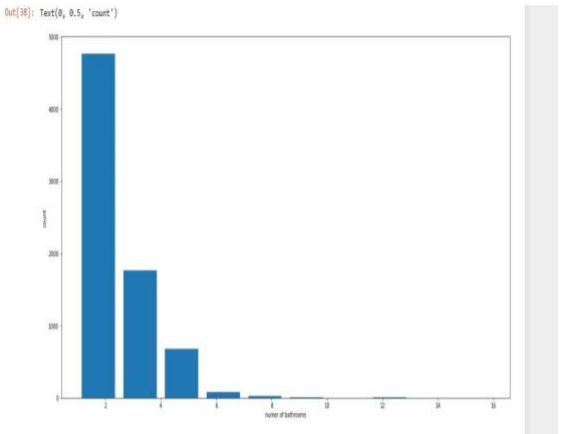
7.1.1 SCATTER PLOTS:





7.1.2. Bar graphs





Outlier removal using Bathroom feature

We can see in the general case there are not 12,13,16 bathrooms in a normal house.

7.1.3 Benefits of Data Science in Real Estate:

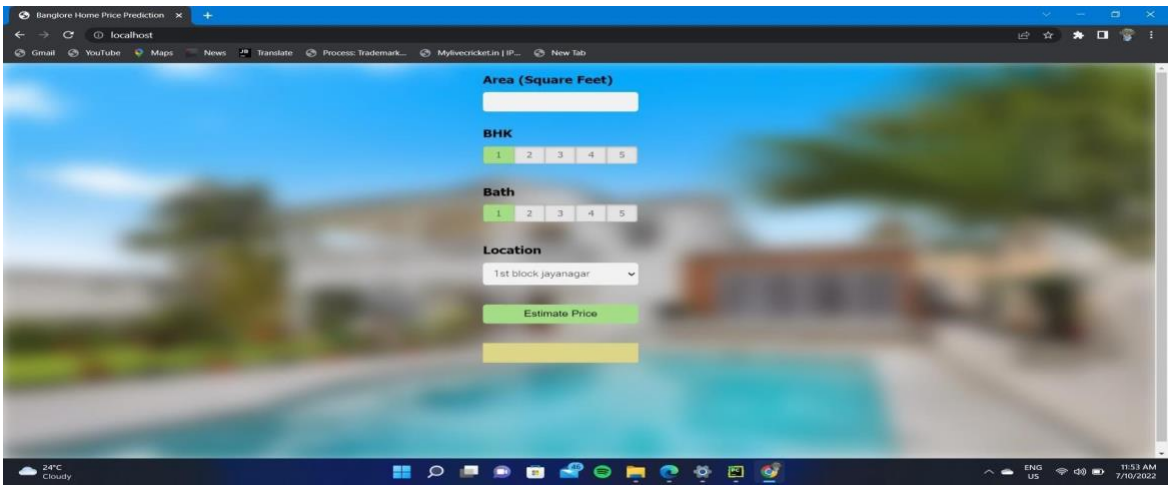
Reduces Risks: with the help of predictive analytics company can use it to estimate the overall condition like its ages, deconstruction history, owner information. the company can provide their customer with up-to-date information so it increases their satisfaction from working with them.

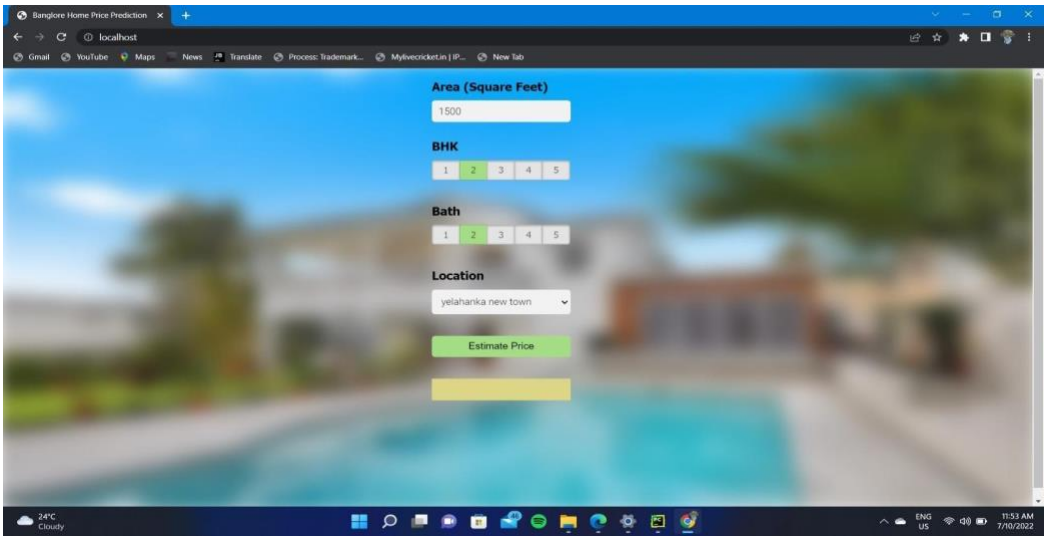
It helps calculate the price: precise cost calculation in real estate is time-consuming, how the Machine learning algorithm can use for the estimate the price of properties with the help of historical data.

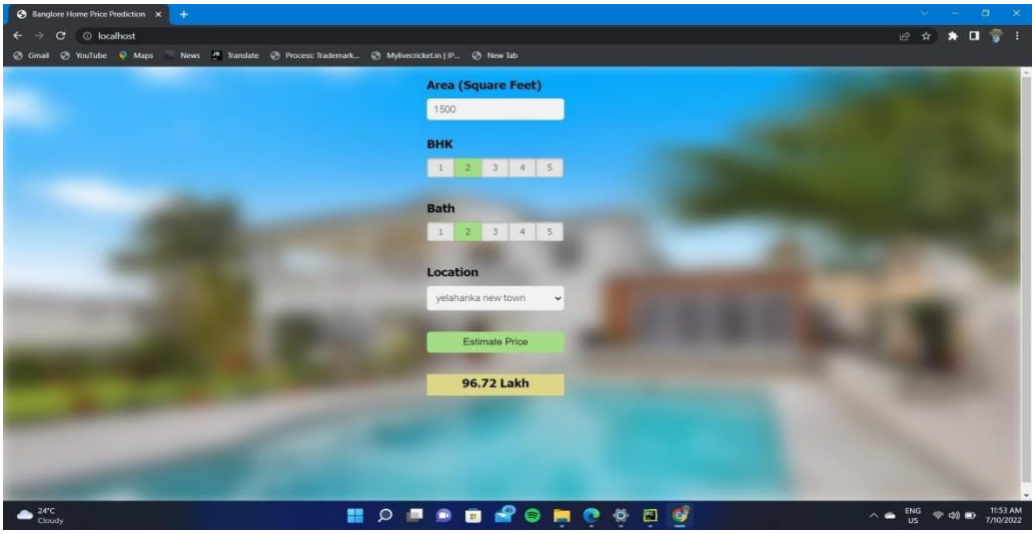
Data-driven decision: Machine learning open many opportunities for the business. just feed the algorithm with data and it will process it to help you make the right decision.

Marketing strategy: with the help of customer information company can plan their future marketing strategy according to customer needs.

7.2. Frontend Snapshots:







8. Expected Outcomes

8.1. Demonstration of working model:

The primary aim of this project is to use these Machine Learning Techniques and curate them into ML models which can then serve the users. The main

objective of a Buyer is to search for their dream house which has all the amenities they need. Furthermore, they look for these houses/Real estates with a price in mind and there is no guarantee that they will get the product for a deserving price and not overpriced. Similarly, a seller looks for a certain number that they can put on the estate as a price tag and this cannot be just a wild guess, lots of research needs to be put to conclude a valuation of a house.

Additionally, there exists a possibility of under-pricing the product. If the price is predicted for these users, this might help them get estates for their deserving prices not more not less.

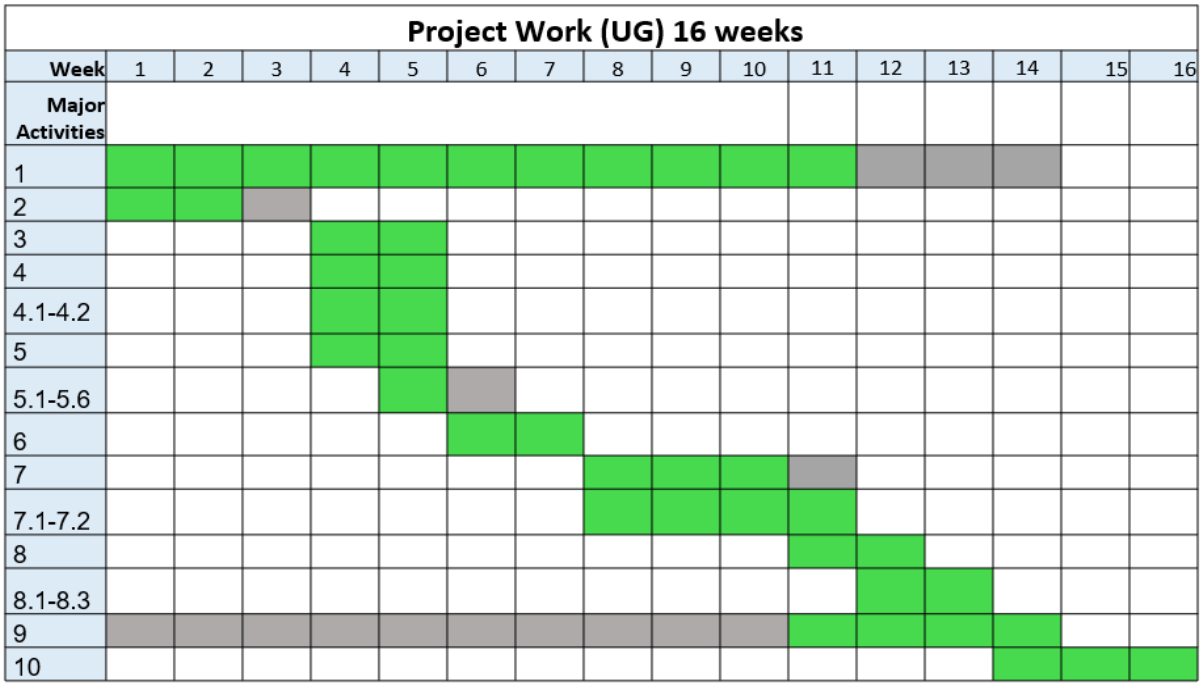
The website also allows user to forecast the predicted house price to a particular date which is also specified by the user. This is done by using another model known as the ARIMA(Auto Regressive Integrated Moving Average Model).

8.2. A new technique

Language Used The “REAL ESTATE PRICE PREDCITION SYSTEM” will be used to for predicting the house price, forecasting that price. This application can be run using website.

We are using Python3 for making machine learning model and Python flask for connectivity and HTML, JavaScript, CSS to develop our web page. We are using anaconda which contains a software Spyder and Jupiter Notebook. Jupiter Notebook contains all updated and latest libraries of python which will be very useful for implementing machine learning model linear regression, ARIMA model and content based Recommendation system. NGINX will be used to deploy the server.

Gantt Chart:



Benefits of Gantt chart:

Here are 6 effective benefits of Gantt Chart:

1. Proper Understanding of Everything
2. One Single View for Everything
3. Easily Break Down Projects into Smaller Pieces
4. Check Dependencies
5. Automatic Scheduler
6. Transparency

References:

1. Real Estate Price Prediction with Regression and Classification, CS 229 Autumn 2016
2. O. Bin, A prediction comparison of housing sales prices by parametric versus semi- parametric regressions, Journal of Housing Economics, 13 (2004) 68-84.
3. T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" In Lecture NotesinComputer.
4. J. Schmidhuber, "Multi-column deep neural networks for image classification," in Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ser. CVPR '12, Washington, DC, USA: IEEE Computer Society, 2012.
5. 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].
6. The elements of statistical learning, Trevor Hastie - Random Forest Generation [8] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
7. https://matplotlib.org/
8. https://www.coursera.org/specializations/recommender-systems
9. https://www.analyticsvidhya.com/blog/2015/08/begi nners-guide-learn-contentbasedrecommendersystems/
10. https://towardsdatascience.com/how-to-buildfrom-scratch-a-content-based-movierecommenderwithnatural-language-processing-25ad400eb243
11. https://www.makaan.com/
12. https://tradingeconomics.com/