Project Report on

Real Estate House Price Prediction Using Linear Regression Model

Submitted by

Y NEERAJA(R170349)

Submitted to

IIIT RK Valley
Idupulapaya,Vempalli,YSR Kadapa
Andhra Pradesh,India PIN 516330.

Under the Guidance of

Ms.M HIMABINDU
Assistant Professor
RGUKT,RK Valley

as a part of
Partial fulfillment of the degree of Bachelor of Technology in
Computer Science and Engineering.

April 2023

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING, RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES,RK VALLEY

CERTIFICATE

This is to certify that the report entitled “**Real Estate House Price Prediction Using Linear Regression Model**” submitted by Y Neeraja (R170349),in partial fulfillment of the requirements for the award of Bachelor of Technology in Computer Science and Engineering is a bonafide work carried out by them under my supervision and guidance.

The report has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

M.HimaBindu,
Project Internal Guide,
Computer Science and Engineering,
R.K Valley, RGUKT.

N.Satyanandaram,
Head of the Departmet,
Computer Science and Engineering,
R.K.Valley, RGUKT.

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant guidance and encouragement crown all the efforts success.

I am extremely grateful to our respected Director,Prof. K. SANDHYA RANI for fostering an excellent academic climate in our institution.

I also express my sincere gratitude to our respected Head of the Department Mr.N SATYANANDARAM for his encouragement, overall guidance in viewing this project a good asset and effort in bringing out this project.

I would like to convey thanks to our guide at college Miss.M HIMABINDHU for her guidance, encouragement, co-operation and kindness during the entire duration of the course and academics.

My sincere thanks to all the members who helped me directly and indirectly in the completion of project work. I express my profound gratitude to all our friends and family members for their encouragement.

TABLE OF CONTENTS

ABSTRACT

In today’s world,real estate is one of the most significant investments,especially in cities like Bangalore,Mumbai,which happens to be a dream city for many people to work and settle in.Therefore,knowing the real-time value of a house is very important before you finance your money on any property.The main purpose of the paper is to predict the market value of the home in Bangalore.Factors like number of bedrooms,Area in squarefoot,number of bathrooms,availability of different types of amentities are taken into account while doing so.

Similarly,consider a situation in which a person needs to sell a house.By using a real estate pricing system,the seller will be able to determine what features he can add to the house so that the house can be sold at a higher price.This prediction is to help a customer look for viable options which are more suited to their requirements.Therefore,in both cases,we can be sure that the home price is good for both the buyer and the seller.Housing prices go up every year, so there is a need for a real estate forecasting system.Estimating the price of a house can help a developer determine the selling price of a house and can help clients set a reasonable time to buy a home.We have used Linear Regression model to predict the cost of the various houses.This model eliminates the need to consult a broker thereby additionally helping the customer.

INTRODUCTION

House price prediction models are a type of machine learning algorithm used to estimate the value of a property based on various factors that affect its value.These models use historical data on real estate sales and their associated attributes to learn patterns and relationships that can be used to predict the price of a property.

The models are typically trained using a dataset that includes information on the features of a property,such as its size,location,number of rooms,and other relevant facors,as well as the selling price of the property.The data is then used to build a predictive model that can estimate the selling price of a property based on its features.

There are several types of models used for house price prediction,including linear regression,decision trees,and neural networks.Each of these models has its strengths and weaknesses and can be used in different scenarios depending on the type and amount of data available.

House price prediction models are useful for a variety of stakeholders in the real estate industry,including buyers,sellers,and real estate agents.By providing accurate predictions of property values,these models can help buyers and sellers make informed decisions about purchasing and selling properties,while real estate agents can use them to provide better advice to their clients.

PURPOSE

The purpose of house price prediction using a linear regression model is to estimate the sale price of a house based on its various features such as location, size, number of bedrooms, number of bathrooms, and other amenities.

Linear regression is a statistical method that enables us to establish a relationship between the independent variables (the features) and the dependent variable (the sale price), and predict the value of the dependent variable for new data points.

By using a linear regression model to predict the sale price of a house, we can help buyers and sellers make informed decisions about pricing and negotiation. Real estate agents and investors can also use this information to evaluate investment opportunities and develop marketing strategies.And also we can analyze the impact of each independent variable on the dependent variable and predict the price of a house with a reasonable degree of accuracy.

It is used to provide valuable insights into the real estate market, help stakeholders make better-informed decisions, and minimize the risk of making a bad investment or missing out an good opportunity.

3.Machine Learning :

Machine learning is a method of data analysis that automates analytical model building.It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.It has wide range of applications including image and speech recognition,predictive analysis and recommendation systems.

3.1Supervised Learning :

Supervised learning is a process of providing input data as well as correct output data to the machine learning model.Tthe aim of a supervised learning is to find a mapping function to map the input variable(x) with the output variable(y).

4.Requirements :

Hardware specifications :

Processor : i3

RAM :4 GB or more

Hard Disk :16 GB or more

GPU : 2 GB

Software specifications :

Platform : Windows operating system

JupyterLab/Visual studio code

python3

5.Libraries

We have imported few Libraries which are needed for the whole process

```
import numpy as np

import pandas as pd

import matplotlib

from matplotlib import pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression
from sklearn.model_selection import ShuffleSplit

from sklearn.model_selection import cross_val_score
```

5.1 Numpy:

Numpy python library is used for including any type of mathematical operation in the code.It is fundamental package for scientific calculation in python.It also supports to add large,multi dimensional arrays and matrices.

5.2 Pandas:

Pandas is an open source library in pythonIt provides ready to use high performance data structures and data analysis tools.Pandas module runs on top of Numpy and it is popularly used for data science and data analytics.

5.3 Matplotlib:

Matplotlib is a python library used to create 2D graphs and plots by using python scripts.It has a module named pyplot which makes things easy for plotting by providing feature to control line styles,font properties,formatting axes etc..

5.4 Scikit-learn:

Scikit-learn(sklearn) is the most useful and robust library for machine learning in python.It provides a selection of efficient tools for machine learning and statistical modelling including classification,regression,clustering and dimensionality reduction via a consistence interface in python.

6. Tools Used :

Visual Studio code :

Visual Studio Code is a free coding editor that helps you start coding quickly. Use it to code in any programming language, without switching editors. Visual Studio Code has support for many languages, including Python, Java, C++, JavaScript, and more.

Visual Studio Code, also commonly referred to as VS Code, is a source-code editor made by Microsoft with the Electron Framework, for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

Python 3 :

Python includes a modular machine learning library known as PyBrain, which provides easy-to-use algorithms for use in machine learning tasks. The best and most reliable coding solutions require a proper structure and tested environment, which is available in the Python frameworks and libraries

Python is consistent and is anchored on simplicity, which makes it most appropriate for machine learning. The Python programming language best fits machine learning due to its independent platform and its popularity in the programming community.

7. Methodology

The process can be divided into several stages which include Data collection,Data cleaning,Feature Engineering and Dimensionality reduction,Outlier detection and Removal,Model building,Model testing,Evaluate the performance of our model.

7.1 Data Collection :

This is the first phase in this process where we collect data from online.The Data set used in this project is downloaded from the **kaggle website,**which is a free source of Data sets for machine learning and Datascience.

It is a reliable source,so we took data from Kaggle.The step of gathering data is the foundation of the machine learning process.

Dataset : Bengaluru_House_Data.csv

Dataset is downloaded from: https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data

The dataset contains total 13320 rows and 9 columns(attributes).Some of the attributes are area_type,availabitlity,location,total_sqft,bathrooms,balcony,price etc,..

Load Bengaluru_House_Data into a Dataframe:

```
import pandas as pd
df1=pd.read_csv("Bengaluru_House_Data.csv")
```

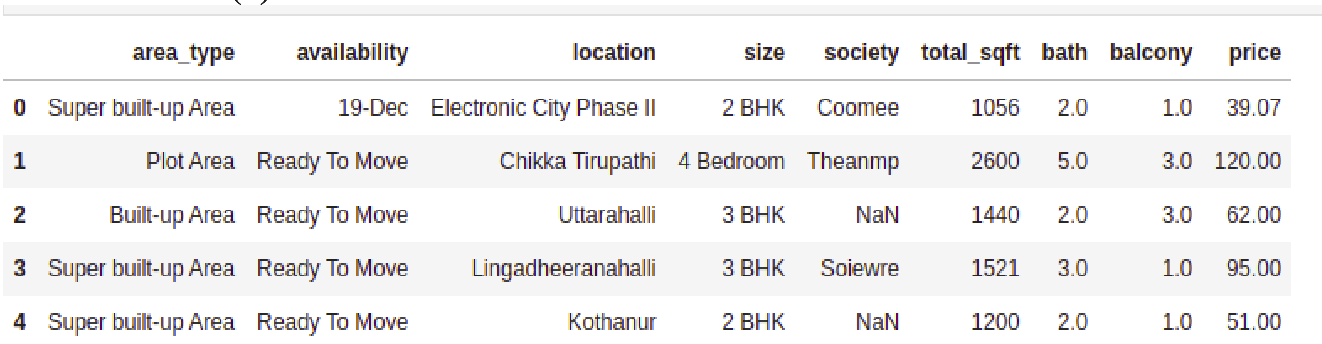
7.2 Exploratory Data Analysis(EDA):

After importing libraries and dataset we will do EDA. It is used to analyze the data using visual techniques.It is used to discover trends,patterns to check assumptions with the help of statistical summary and graphical representations.

7.2.1 Know about the data :

View top 5 rows

df1.head(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

Identify the variables and its datatypes

df1.shape - which gives (13320,9)

The following features that are required to predict the price :

location: It gives the area at which the house is located
size:It refers to the size of house whether its 2bhk or 3bhk and so on
total_sqft:It refers to the total square foot area of the house
bath:It refers to number of bathrooms the house contain

By observing these, we don’t require the columns include availability, balcony, society, area_type.so we will drop the features that are not required to build our model.

df2 = df1**.**drop(['area_type','society','balcony','availability'],axis='columns')
df2.shape -which gives value(13320,5)

7.2.2 Data Cleaning :

Data cleaning will handle the missing values.Missing data in the dataset can reduce the fit of a model or can lead to biased model because we have not analysed the behaviour and relationship with other variables correctly.It can lead to wrong prediction or classification.

By analysing the details of dataset,it is found that few features have null values and bath variable has maximum of 73 null values.

df3 = df2**.**dropna()

After dropping null values,we have zero missing values.Check it by using isnull() method which results in zero null values.

7.2.3 Feature Engineering :

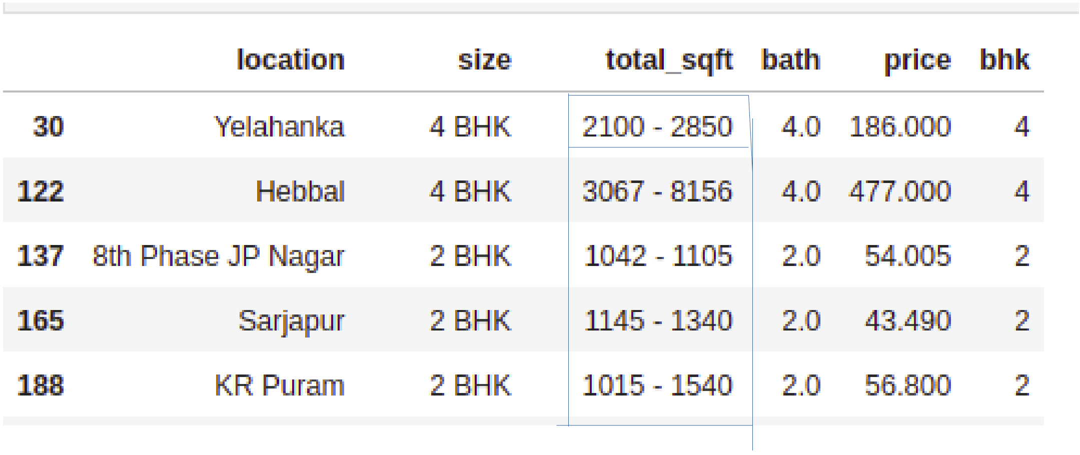
Feature Engineering is the pre-processing step of machine learning,which extracts features from raw data.It helps to represent an underlying problem to predictive models to a better way.In this data set size column contains different values like 2bhk,2 bedroom which is same creates problem.To avoid this we will create a new column called bhk which contains the integer represents number of bedrooms.

- **Size:**

```
df3['bhk'] = df3['size'].apply(lambda x: int(x.split(' ')[0]))
```

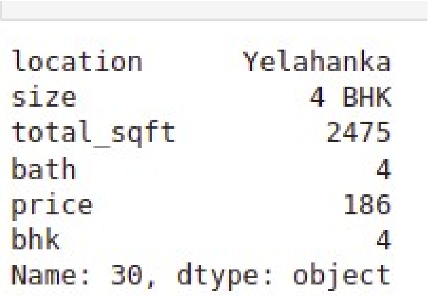
- **total_sqft:**

And also we found that total_sqft contain range of values other than the single value,that need to be converted to single value.



	location	size	total_sqft	bath	price	bhk
30	Yelahanka	4 BHK	2100 - 2850	4.0	186.000	4
122	Hebbal	4 BHK	3067 - 8156	4.0	477.000	4
137	8th Phase JP Nagar	2 BHK	1042 - 1105	2.0	54.005	2
165	Sarjapur	2 BHK	1145 - 1340	2.0	43.490	2
188	KR Puram	2 BHK	1015 - 1540	2.0	56.800	2

```
def convert_sqft_to_num(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_to_num)
df4.loc(30)
```

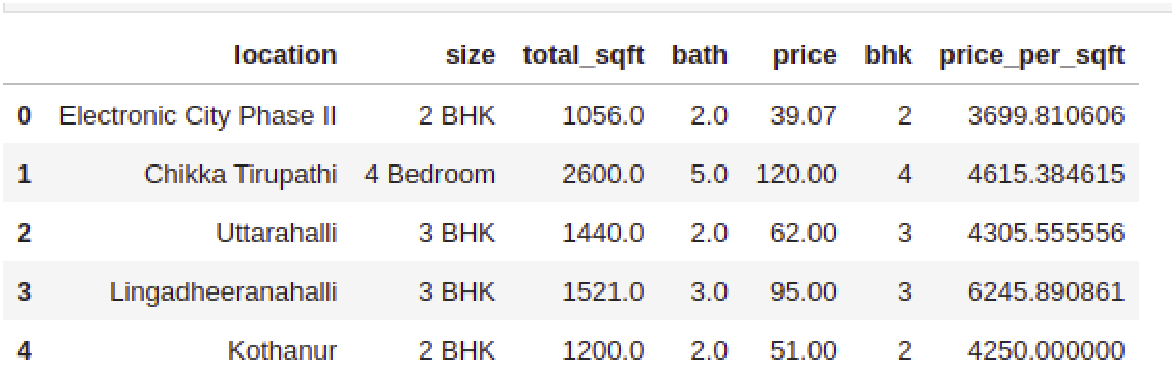


```
location      Yelahanka
size              4 BHK
total_sqft         2475
bath                  4
price               186
bhk                   4
Name: 30, dtype: object
```

- **new feature – price_per_sqft**

We are adding a new feature called price_per_sqft which is important and will help us to clean the outliers.we create it by using the price and total_sqft (price/total_sqft)

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



	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699.810606
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615.384615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305.555556
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245.890861
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250.000000

7.2.4

Dimensionality Reduction

Dimensionality reduction technique can be defined as, “It is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information”.Any location having less than 10 data points should be tagged as ‘other’ location.By using this number of categories can be reduced by huge amount.We are having 1287 unique locations before applying Dimensionality reduction.After doing this we are having only 241 locations as unique which is much better.

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

7.2.5 Outlier Removal

- **Using Business Logic**

As a data scientist when you have a conversation with your business manager(expert in real estate) will tell you that normally square ft per bedroom is 300(i.e. 2 bhk apartment is minimum 600 sqft.If you have 400 sqft apartment with 2 bhk that seems to be suspicious and can be removed as an outlier.We will remove such outlier by keeping our minimum threshold per bhk to be 300 sqft.

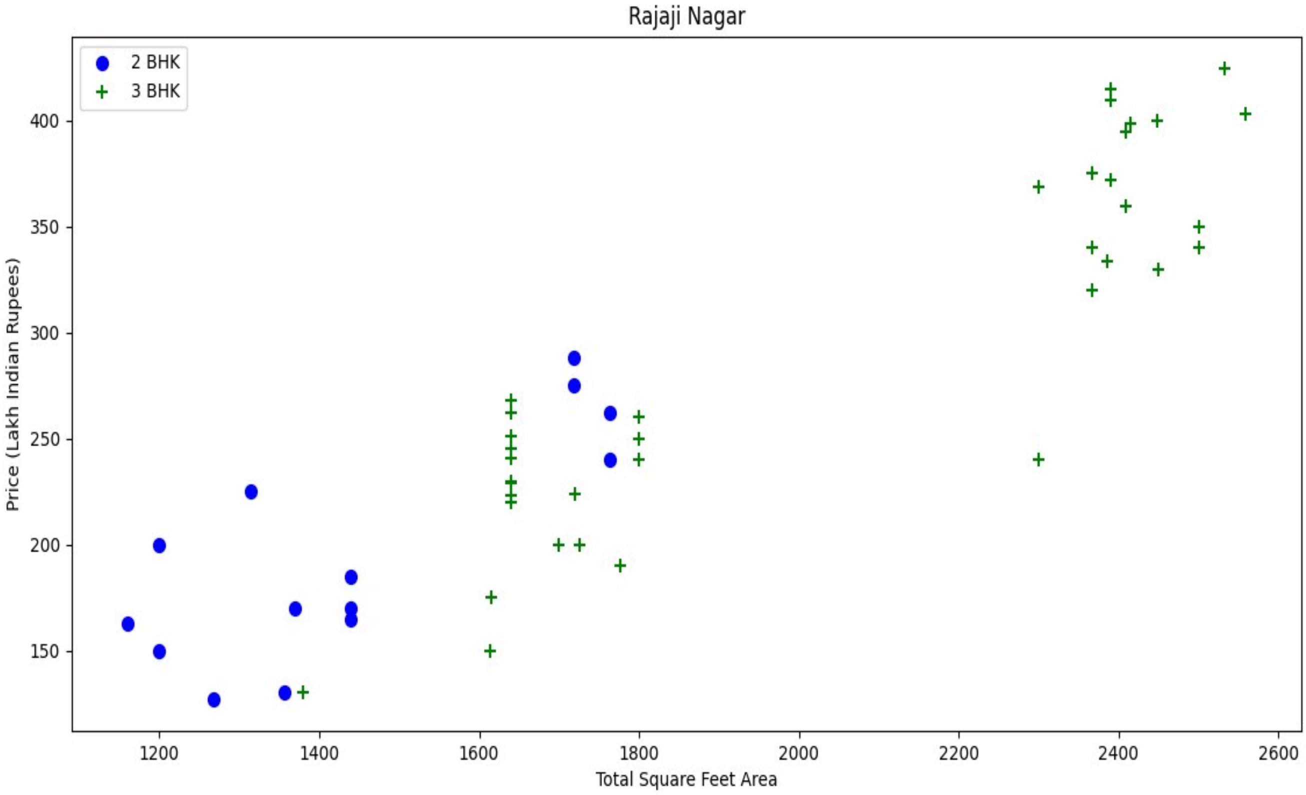
```
df6=df5[~(df5.total_sqft/df5.bhk<300)]
```

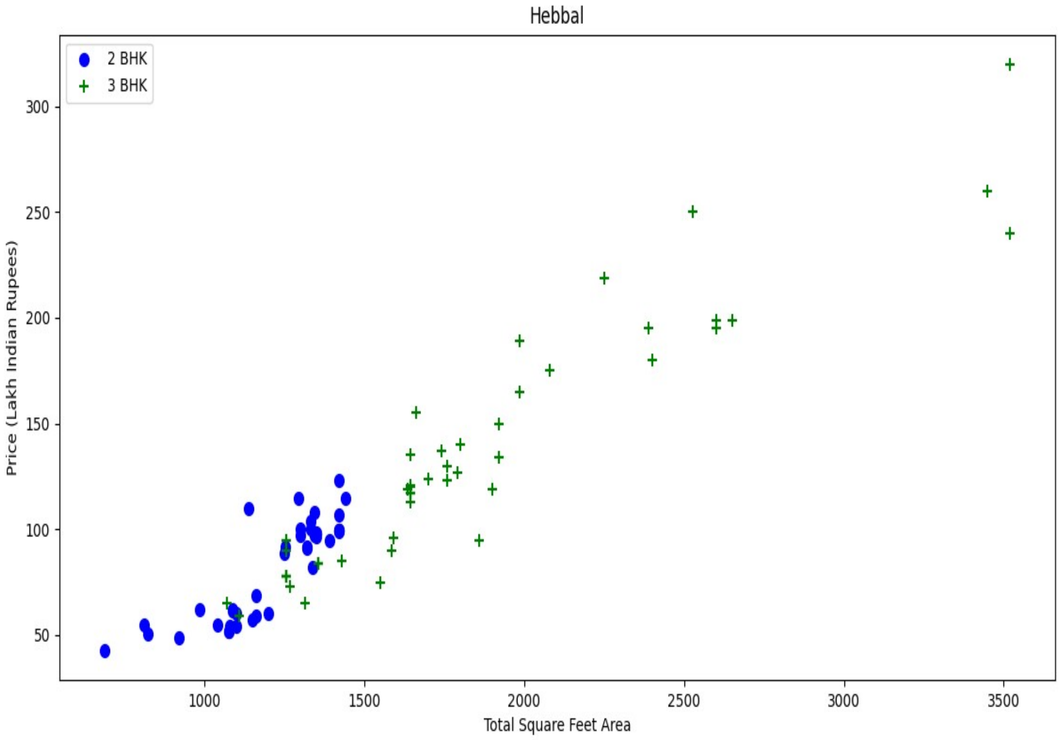
This results in removal of around 740 data points.

- **Using Standard deviation and Mean**

Here we find that min price per sqft is 267 rs/sqft where as 1200000,this shows a wide variation in property prices.We should remove outliers per location using mean and one standard deviation.We keep records of house whose price per sqft is greater than difference between mean and standard deviation and is less than the sum of mean and standard deviation.

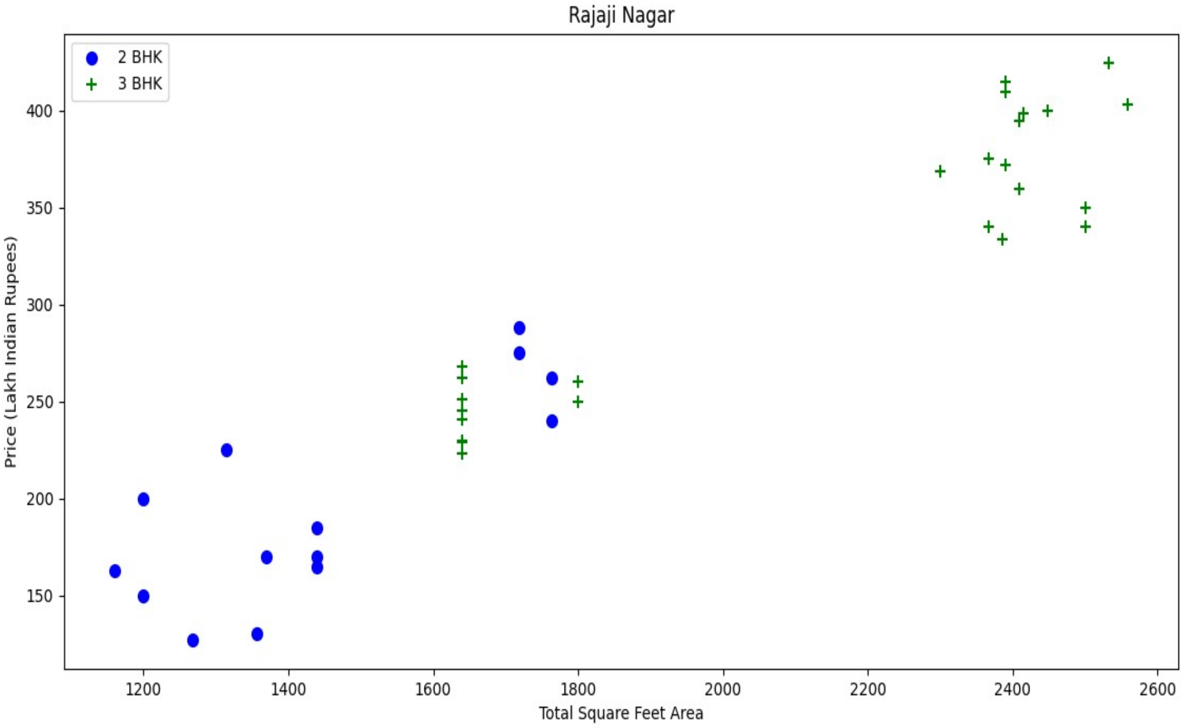
For a given location the 2bhk and 3bhk property prices look like



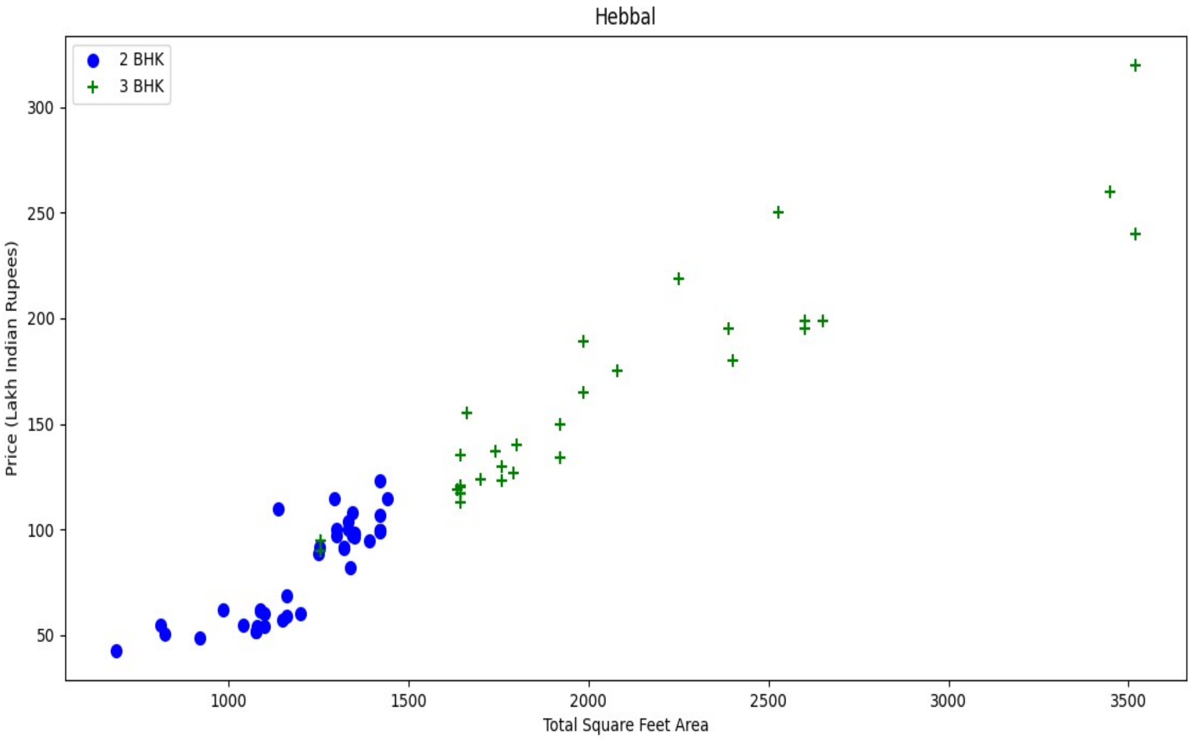


We should remove the properties where for same location,the price of (for example) 3 bedroom apartment is less than 2 bedroom apartment(with same square ft area).

After outlier removal:Rajaji Nagar

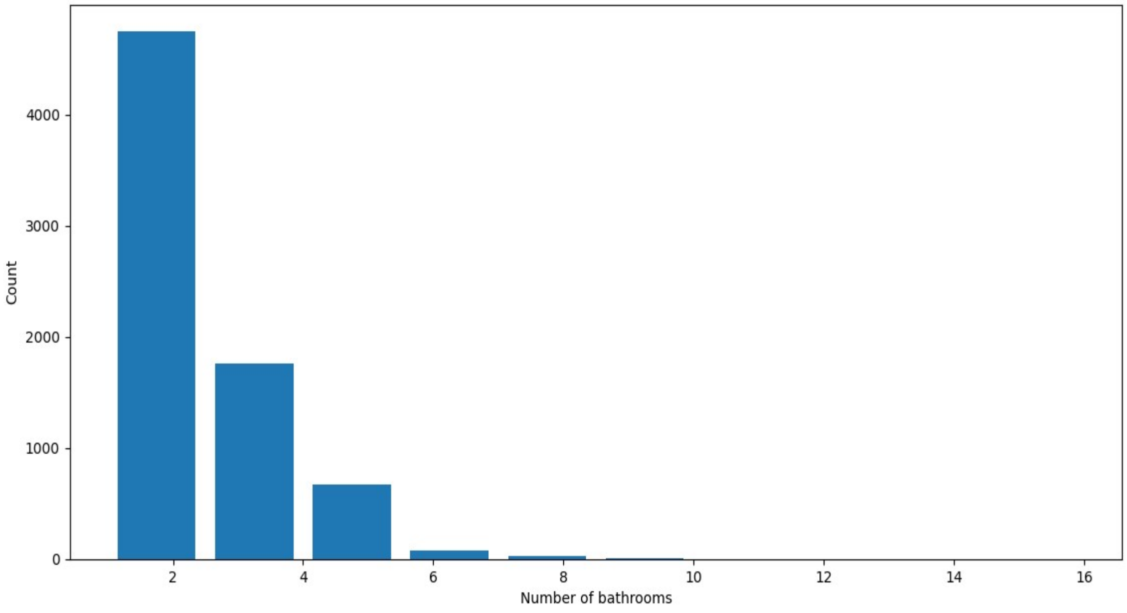


After outlier removal:Hebbal



- **Outlier Removal using Bathroom feature:**

  no.of bathrooms vs count :



By observing above chart we can say that 2 to 5 bathrooms have the most count in the data set.It is unusual to have 2 more bathrooms than number of bedrooms in a home.It is enough to have that,so will remove those outliers

```
df9 = df8[df8.bath<df8.bhk+2]
print(df9.shape)
=>(7239,7)
```

7.2.6 One Hot Encoding:

One hot encoding is a technique that we use to represent categorical variables as numerical values in a machine learning model.It can help to avoid the problem of ordinality, which can occur when a categorical variable has a natural ordering.

```
encode=pd.get_dummies(df10.location)
```

```
df11=pd.concat([df10,dummies.drop(‘other’,axis=’columns’)],axis=’columns’)
df12=df11.drop(‘location’,axis=’columns’)
df12.shape
=>(7239,244)
```

7.3 Model Building :

Prediction using Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

```
X = df12.drop(['price'],axis='columns')
y = df12.price
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)
print(lr_clf.score(X_test,y_test))
```

K Fold cross validation:

Use K Fold cross validation to measure accuracy of our LinearRegression model

```
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)
print(cross_val_score(LinearRegression(), X, y, cv=cv))
```

We can see that in 5 iterations,we got a score above 80% all the time.This is pretty good but we want to test few other algorithms for regression to see if we can get even better score.

Linear Regression :

```
lr_clf=LinearRegression()
lr_clf.fit(X_train,y_train)
lr_clf.score(X_test,y_test)
```

Lasso Regression :

```
lasso=Lasso()
lasso.fit(X_train,y_train)
lasso.score(X_test,y_test)
```

Decision Tree Regression :

```
dtr=DecisionTreeRegressor()
dtr.fit( X_train,y_train)
dtr.score(X_test,y_test)
```

```
Output:0.847796
       0.726738
       0.716064
```

8.Observations :

By observing the above results we say that Linear Regression gives the best score of all the models.Lasso Regression gives 72.67% , Decision Tree Regression gives 71.61% and Linear Regression gives 84.78% accuracy.So we will say that the Linear Regression model suits the best among all.

The price mainly depends on the Location,size,total square feet,number of bathrooms,number of bedrooms.We can also use GridSearchCV method to find the accurate model for house price pediction.

Therefore we can say Linear Regression is accurate model among all.

Accuracy of the model is : 0.847796

9.Analysis and Design :

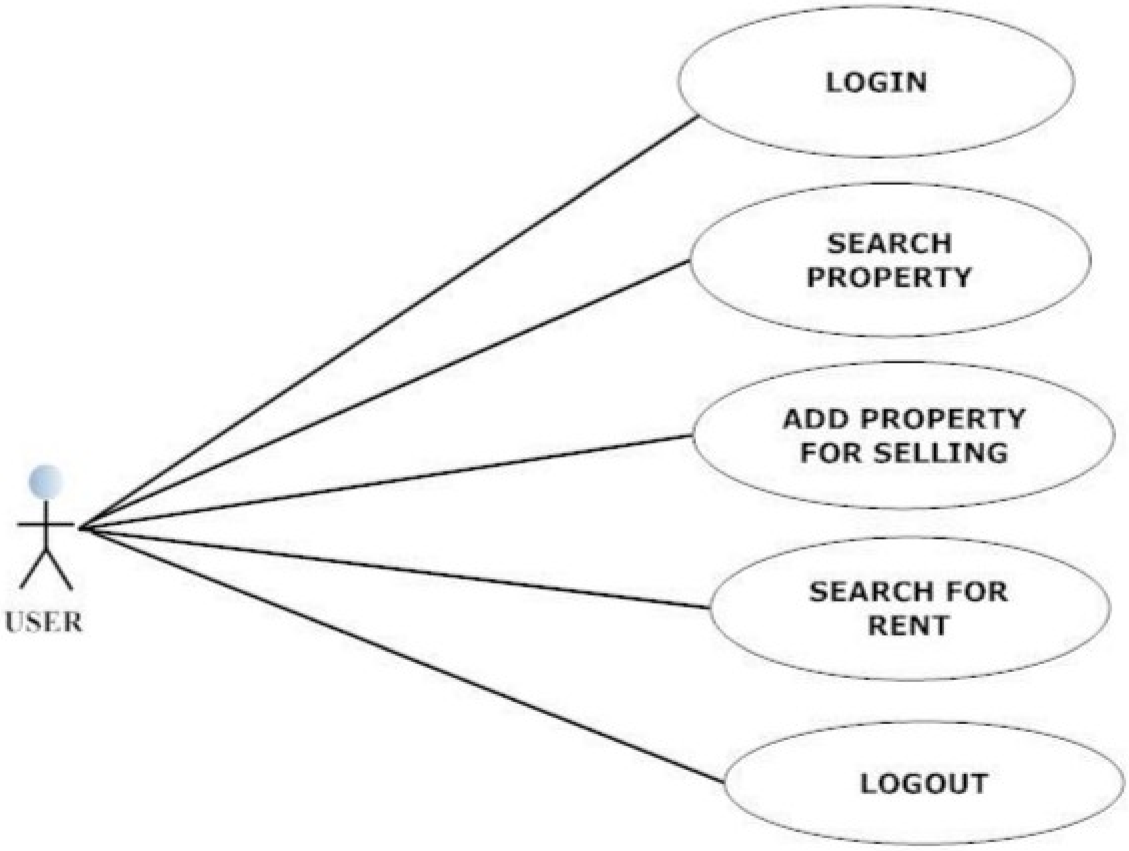
9.1 Use case Diagram

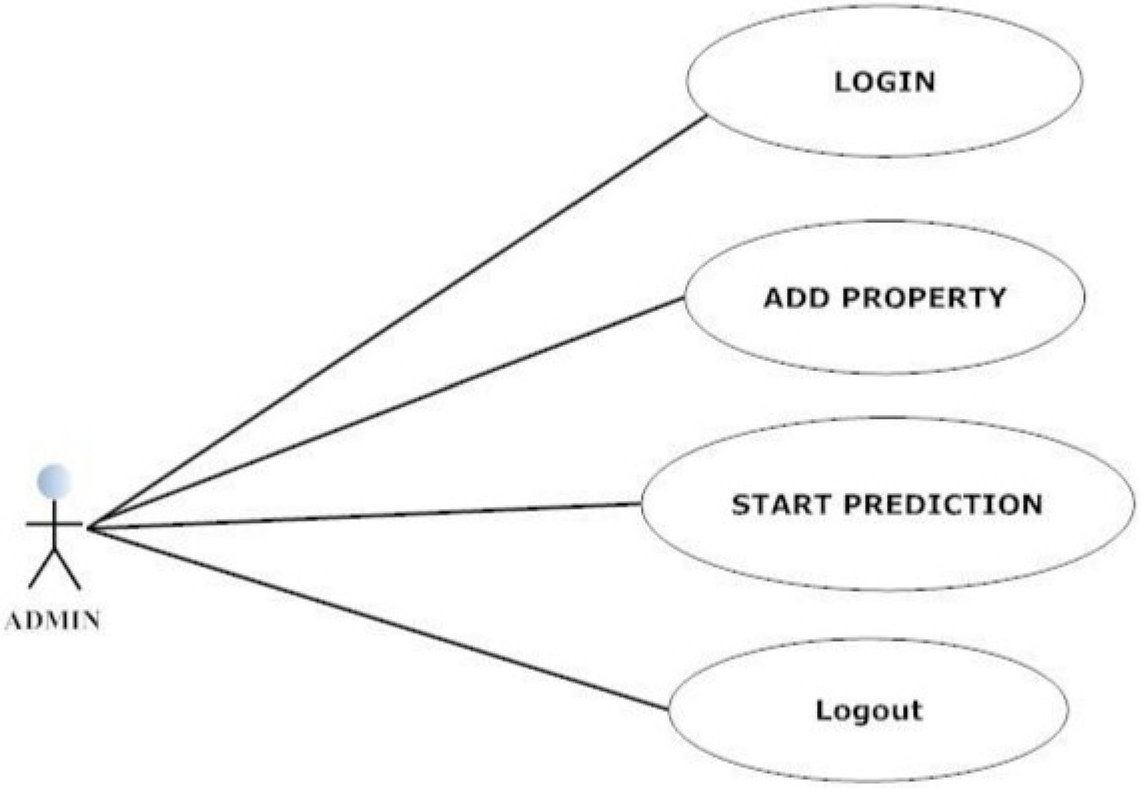
Use case diagrams model behavior within a system and helps the developers understand of what the user require. The stick man represents what’s called an actor.

Use case diagram can be useful for getting an overall view of the system and clarifying that can do and more importantly what they can’t do.

Use case diagram consists of use cases and actors and shows the interaction between the use case and actors.

- The purpose is to show the interactions between the use case and actor.
- To represent the system requirements from user’s perspective.
- An actor could be the end-user of the system or an external system.

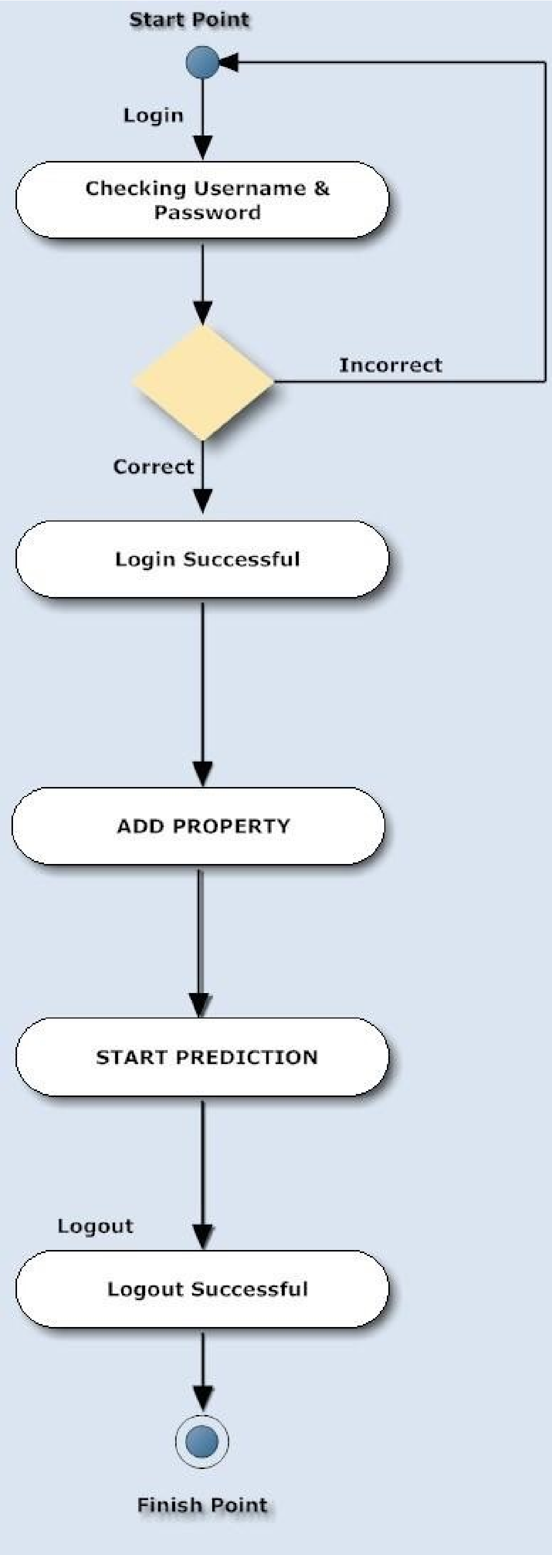




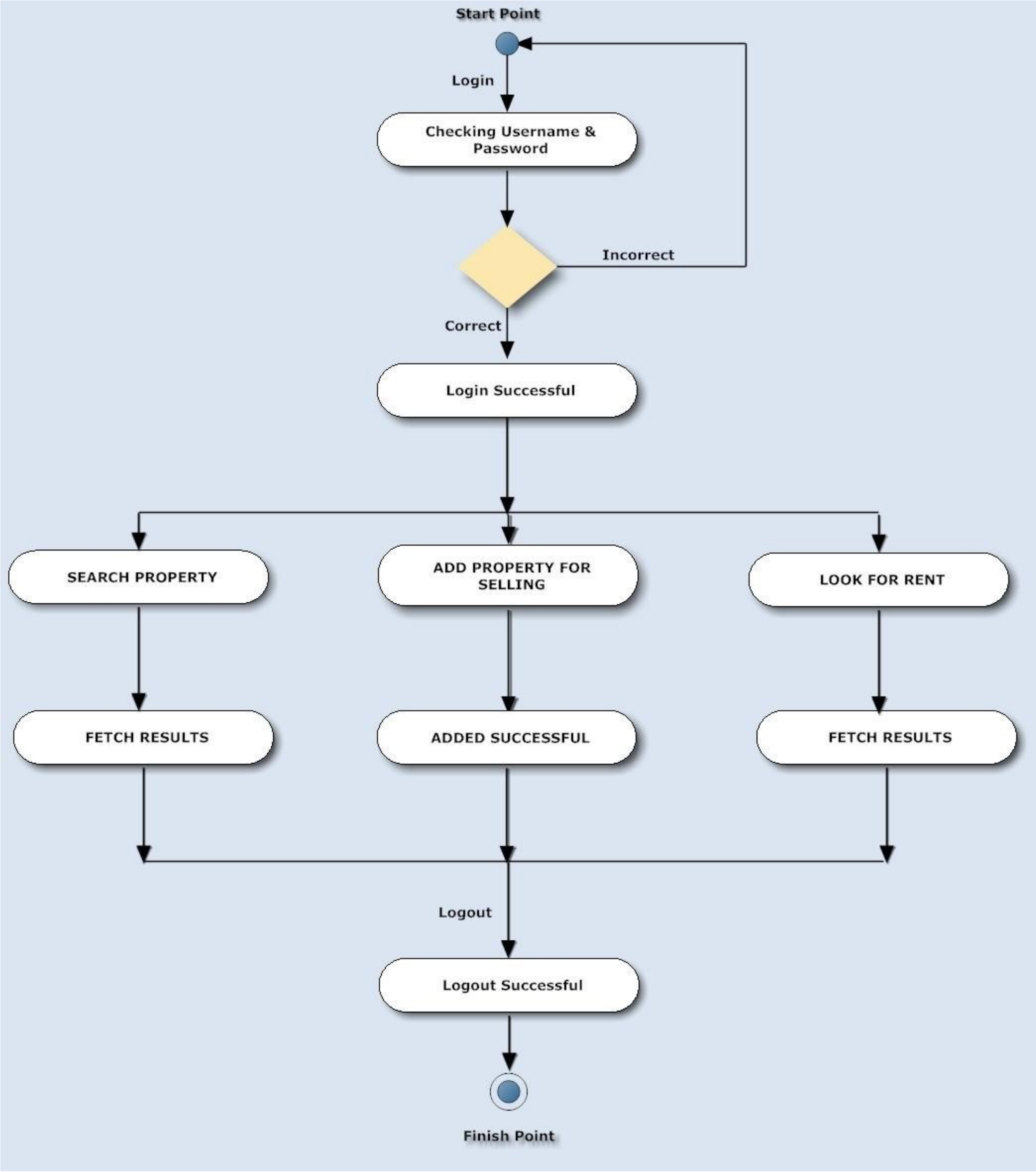
9.2Activity Diagram :

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. This flow can be sequential, branched, or concurrent. Activity diagrams deal with all type of flow control by using different elements such as fork, join, etc.

Admin activity: -Admin can log in by providing his user name and password if it is incorrect, screen will show invalid message if it is correct log in successful message will be shown Admin can add his properties and start predictions over it.


Start Point
Login
Checking Username & Password
Incorrect
Correct
Login Successful
ADD PROPERTY
START PREDICTION
Logout
Logout Successful
Finish Point

User activity:-User can Search property, add property for selling and can also look for rent by search method after logging in. Results are fetched and added from the main database.



10. Conclusion :

In this project, the website allows the user to give property details according to his/her requirement. The system makes optimal use of the Data mining Algorithm i.e Linear Regression. The Linear Regression algorithm is used to predict the house price according to the property requirement given by the customer with accuracy of 84.78%.This system will help the user to get the best and relevant real estates residential properties according to the budget given by the user. The main objectiveof using this prediction, forecasting and recommendation system is to reduce the human physical calculation, time and carry out the whole process at ease