# "House Price Prediction Using Deep Learning"

## A Project Report Submitted to the

## University of Mumbai

## Mumbai In partial fulfilment of the course work

## leading to Semester VII

By


| Name | Roll No. |
|---|---|
| Swapnil Patil | 49 |
| Ruturaj Raut | 55 |
| Avdhut Shinde | 60 |


**Under the guidance of**

**Prof. Shashikant Patil**



**Department of**

**Computer Engineering**

**Vishwaniketan's**

**Institute of Management Entrepreneurship & Engineering**

**Technology [iMEET]**


**[2020-21]**

s

**Vishwaniketan's**

**Institute of Management Entrepreneurship & Engineering Technology**

**2020-2021**


# CERTIFICATE

This is to certify that the project report titled, **"House Price Prediction using Deep Learning"**, duly submitted by the following students-

| Name | Roll No. |
|---|---|
| Swapnil Patil | 49 |
| Ruturaj Raut | 55 |
| Avdhut Shinde | 60 |

Has been completed under my supervision in a satisfactory manner in a partial fulfillment of the requirements for the award of semester VII Bachelor's Degree in **Computer Engineering** to be conferred by the **University of Mumbai**. In my opinion, the work embodied in this report is comprehensive and fit for evaluation.

**Prof. Abhijeet Patil**
**Project Guide and HOD**

**Dr. B. R. Patil**
**Principal**

# PROJECT REPORT APPROVAL SHEET

The Project Report Titled "**Simulating Customized Game-play using ML Agents**" submitted by the students

| Name | Roll No. |
|---|---|
| Swapnil Patil | 49 |
| Ruturaj Raut | 55 |
| Avdhut Shinde | 60 |

Is examined by the board of examiners and approved for further perusal

Sign: ----------

Name: ----------

(Examiner-I)

Sign: ----------

Name: ----------

(Examiner-II)

Date:

Place:

# Declaration

We declare that this written submission represents our ideas in our own words and where others ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

| | | |
|---|---|---|
| Swapnil Patil | Ruturaj Raut | Avdhut Shinde |

Date:

# 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 analyze 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.

The accuracy of the prediction is evaluated by checking the root square and root mean square error scores of the training model. The test is performed after applying the required pre-processing methods and splitting the data into two parts. However, one part will be used in the training and the other in the test phase. We have also presented a binning strategy that improved the accuracy of the models.


## Table of Contents

# Chapter 1
# Introduction

## 1.1 Introduction

1. Buying a house is stressful thing.
2. Buyers are generally not aware of factors that influence the house prices.
3. Hence real estate agents are trusted with the communication between buyers and sellers as well as lying down a legal contract for the transfer. This just creates a middleman and increases the prices of houses.
4. The problem falls under the category of supervised learning algorithms. The dataset we'll be using is the Metropolitan areas Dataset. The dataset comprises 7 input features and one target feature. The input features include features that may or may not impact the price.

## 1.2 Objective

- 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.
- The functioning involves a website which accepts customers specifications and then combines the application of Naive bayes algorithm of data mining.
- This application will help customers to invest in an estate without approaching an agent. It also decreases the risk involved in the transaction.
- The current property buying or selling is hectic and expensive. As the customer has to roam places and has to pay commission to the Real estate agent.
- Also, the customer/buyer does not know whether the property is profitable in future or not. Hence, we design a website using data mining techniques to overcome the drawbacks of current system as everything is web based.

## 1.3 Organization of the report

Chapter 1: This gives the Introduction to the topic and the need of the project highlighting the main objectives and the scope of the project.
Chapter 2: It is the Literature Survey which gives a brief description of the similar works performed in the same domain or investigation. It presents a critical appraisal of the previous work published in the literature pertaining to the topic of the investigation.

Chapter 3: Gives a brief explanation about the existing system and proposed model along with the requirements for the project.

Chapter 4: This Chapter includes and Explains the Technology used along with a brief implementation of the Project.

Chapter 5: It is the chapter that includes a thorough evaluation of the investigation carried out and bring out the contributions from the study. It mentions the Conclusion regarding the implementation details and the analysis of the performance measures.

Chapter 6: the References used for the entire project are mentioned in this Chapter. Chapter 7:

Acknowledgements are mentioned in this chapter.

# Chapter 2

# Literature Review

## 2.1 Literature Review

- Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh had proposed an advanced house prediction system using linear regression.

- This system's aim was to make a model that can give us a good house price prediction based on other variables.

- They used the Linear Regression for Ames dataset and hence it gave good accuracy.

- The house price prediction project had two modules namely, Admin and the User.

- Admin can add location and view the location. Admin had the authority to add density on the basis of per unit area.

- Users can view the location and see the predicted housing price for that particular location.

- Machine learning has many application's out of which one of the applications is prediction of real estate. The real estate market is one of the most competitive in terms of pricing and same tends to be vary significantly based on lots of factor, forecasting property price is an important modules in decision making for both the buyers and investors in supporting budget allocation.

- finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. The study on land price trend is felt important to support the decisions in urban planning. The real estate system is an

- unstable stochastic process.

- Investors decisions are based on the market trends to reap maximum returns. Developers are interested to know the future trends for their decision making. To accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis,

modelling and forecasting.

## 2.2 Literature Summary

| Sr. no | Research Paper Tile | Year and Publication | Technology | Advantages | Disadvantages |
|---|---|---|---|---|---|
| 1 | US housing price | 2007, Rapach Furthermore strauss (2007) use an auto regressive dispersed slack (ARDL) model | - | looking at state, territorial and national level variables. | Review more than implementation. Presents an idea of how a structure can be implemented. |
| 2 | Machine Learning based Predicting House Prices using Regression Techniques | 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) | - | Based only on monthly real house price growth rates.<br>Future direction of real house prices. | Narrows down the research to a very specific genre. Architecture breakdown of the elements of the game was rigid and fixed. |
| 3 | House Price Prediction Using Machine Learning and Neural Networks | 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) | ML-Algo , Regression technique. | We use various regression techniques in this pathway.<br>The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied | It needs a lot of data and a lot of computation. It is data-hungry. Reinforcement learning assumes the world is Markovian, which it is not. |

# Chapter 3

# Project Implementation

## 3.1 Project Implementation

### Existing Methodology and System

1. Problem Formulation -
    - Convert your business problem into statistical problem. (attrition or acquisition)
    - Clearly define dependent and independent variable.
    - Identify what you want to predict.
2. Data Cleaning –
    - Transform collected data into usable data table format.
3. Data Pre-Processing-
    - Filter Data (Remove unwanted data)
    - Aggregate Value (Calculate sum, avg, ratio, etc..)
    - Handling blank value in dataset.
    - Outlier Treatment (Handling unacceptably large or small value of variable)
    - Variable Transformation (Make mathematical transformation i.e log or sin)
    - Variable Reduction (Remove less useful or harmful variable)
    -
4. Model Training –
    - Test-Train-Split [to split data in three different format]
    - Output = Input function.
5. Performance Metrics and Validation –
    - In sample error- error resulted from applying your prediction algorithm to dataset you built it with.
    - Out of sample error – error resulted from applying your prediction algorithm to new dataset.
6. Prediction –
    - Setup pipeline to use your model in real life.
    - Improve by monitoring your model over time.
    - Try to automate.

## 3.2 Proposed Methodology and System

1. Early Stop Model
2. Training Model
3. Neural Network Model
4. Class hierarchy
5. Exhibition Flow

## 3.3 Requirements

### 3.3.1 Hardware Requirement

- **Processor: Intel i3/i5**
- **Memory: Minimum 4GB RAM**
- **Disk Space: 100 MB**
- **Internet connectivity**

### 3.3.2 Software Requirement

- **Operating System:** Any Operating System (OS) which can support Internet browser preferred support for Google Chrome.
- **Anaconda –** Computer Program
- **Jupyter Notebook**
- **Pycharm**
- **Tensarflow**
- **UI – (Django , Flask )**

### 3.3.3 Libraries Used for this Project include –

- **Pandas**
- **NumPy**
- **Matplotlib**
- **Seaborn**
- **Scikit Learn**
- **XG Boost**

# Chapter 4

# Technology Used and Implementation

## 4.1 Dataset

Here we have web scrapped the Data from Kaggle.com website which is one of the leading real estate websites operating in INDIA.

## Dataset looks as follows-

| | Price | PricePerSqft | Area_Sqm | Location | Bedrooms | Latitude | Longitude | PricePerSqM |
|---|---|---|---|---|---|---|---|---|
| 0 | 13300000 | 16625 | 74.32 | Kandivali (East) | 2 | 19.210200 | 72.864891 | 178885.00 |
| 1 | 9000000 | 15666 | 55.74 | Ramgad Nagar | 1 | 19.167700 | 72.949300 | 168566.16 |
| 2 | 9000000 | 19148 | 43.66 | Mahakali Caves | 1 | 19.130609 | 72.873816 | 206032.48 |
| 3 | 9000000 | 10588 | 78.97 | Louis Wadi | 2 | 19.126005 | 72.825052 | 113926.88 |
| 4 | 100000000 | 20000 | 464.51 | Barrister Nath Pai Nagar | 5 | 19.075014 | 72.907571 | 215200.00 |

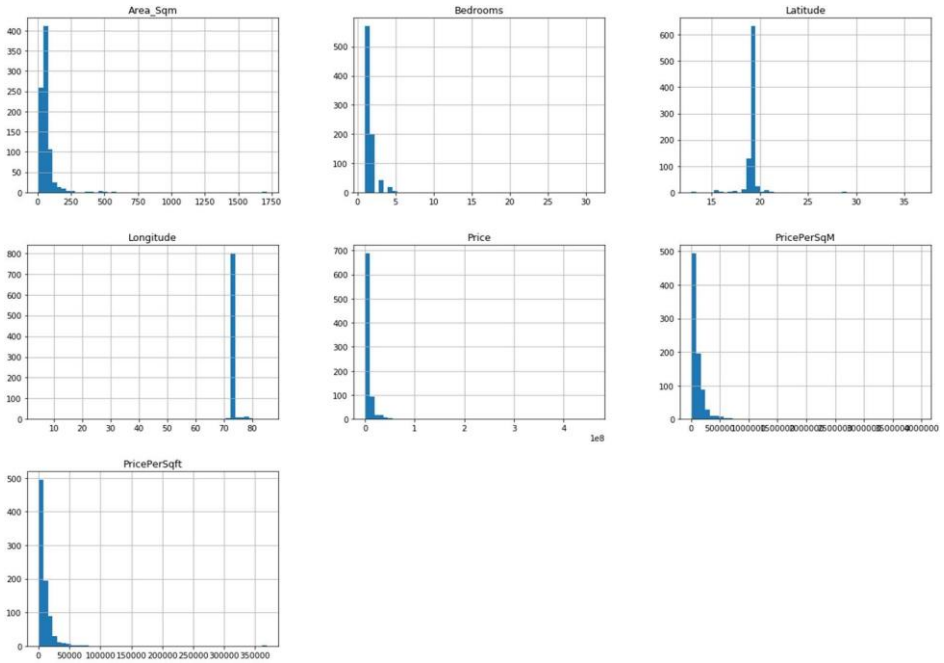
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 840 entries, 0 to 839
Data columns (total 6 columns):
Price          840 non-null int64
Area_Sqm       840 non-null float64
Bedrooms       840 non-null int64
Latitude       840 non-null float64
Longitude      840 non-null float64
PricePerSqM    840 non-null float64
dtypes: float64(4), int64(2)
memory usage: 39.5 KB
```

## 4.2 Data Exploration

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python**.** Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

## 4.3 Data Visualization

visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions.



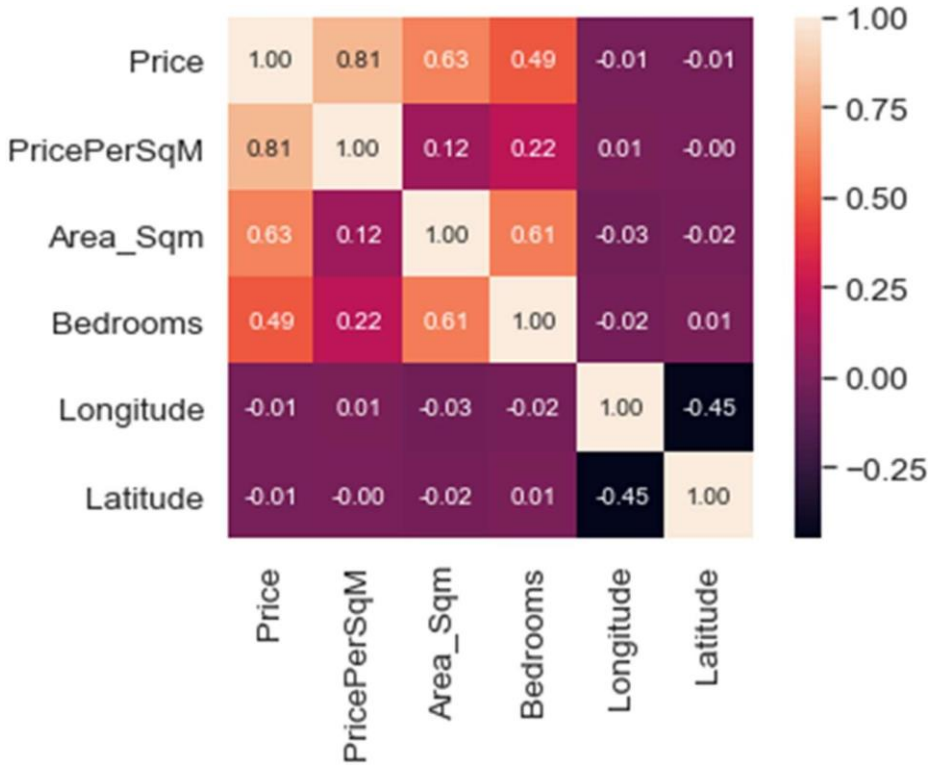
## 4.4 **Data Selection**

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity.

The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

| | Price | Area_Sqm | Bedrooms | Latitude | Longitude | PricePerSqM |
|---|---|---|---|---|---|---|
| 0 | 13300000 | 74.32 | 2 | 19.210200 | 72.864891 | 178885.00 |
| 1 | 9000000 | 55.74 | 1 | 19.167700 | 72.949300 | 168566.16 |
| 2 | 9000000 | 43.66 | 1 | 19.130609 | 72.873816 | 206032.48 |
| 3 | 9000000 | 78.97 | 2 | 19.126005 | 72.825052 | 113926.88 |
| 4 | 100000000 | 464.51 | 5 | 19.075014 | 72.907571 | 215200.00 |

**Correlation Heatmap**



# 4.4 Implementation

## Dataset:

In [2]:
```
df1 = pd.read_csv("E:/HousePricePredictionMumbai/Mumbai.csv")
df1.head(5)
```

Out[2]:

| | Price | Area | Location | size(rk or bhk) | Num_of_bedrooms | Resale | MaintenanceStaff | Gymnasium | SwimmingPool | LandscapedGardens | ... | LiftAvailable | BED | Vaa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 4850000 | 720 | Kharghar | bhk | 1 | 1 | 1 | 0 | 0 | 0 | ... | 1 | 0 | |
| 1 | 4500000 | 600 | Kharghar | rk | 0 | 1 | 1 | 1 | 1 | 0 | ... | 1 | 0 | |
| 2 | 6700000 | 650 | Kharghar | rk | 0 | 1 | 1 | 1 | 1 | 0 | ... | 1 | 0 | |
| 3 | 4500000 | 650 | Kharghar | rk | 0 | 1 | 1 | 0 | 0 | 1 | ... | 1 | 1 | |
| 4 | 5000000 | 665 | Kharghar | rk | 0 | 1 | 1 | 0 | 0 | 1 | ... | 1 | 0 | |

5 rows × 41 columns

In [4]:
```
#drop unused columns
df2 = df1.drop(['Resale','MaintenanceStaff','Gymnasium','SwimmingPool','LandscapedGardens','JoggingTrack','RainWaterHarvesting',
df2.head(8)
```

Out[4]:

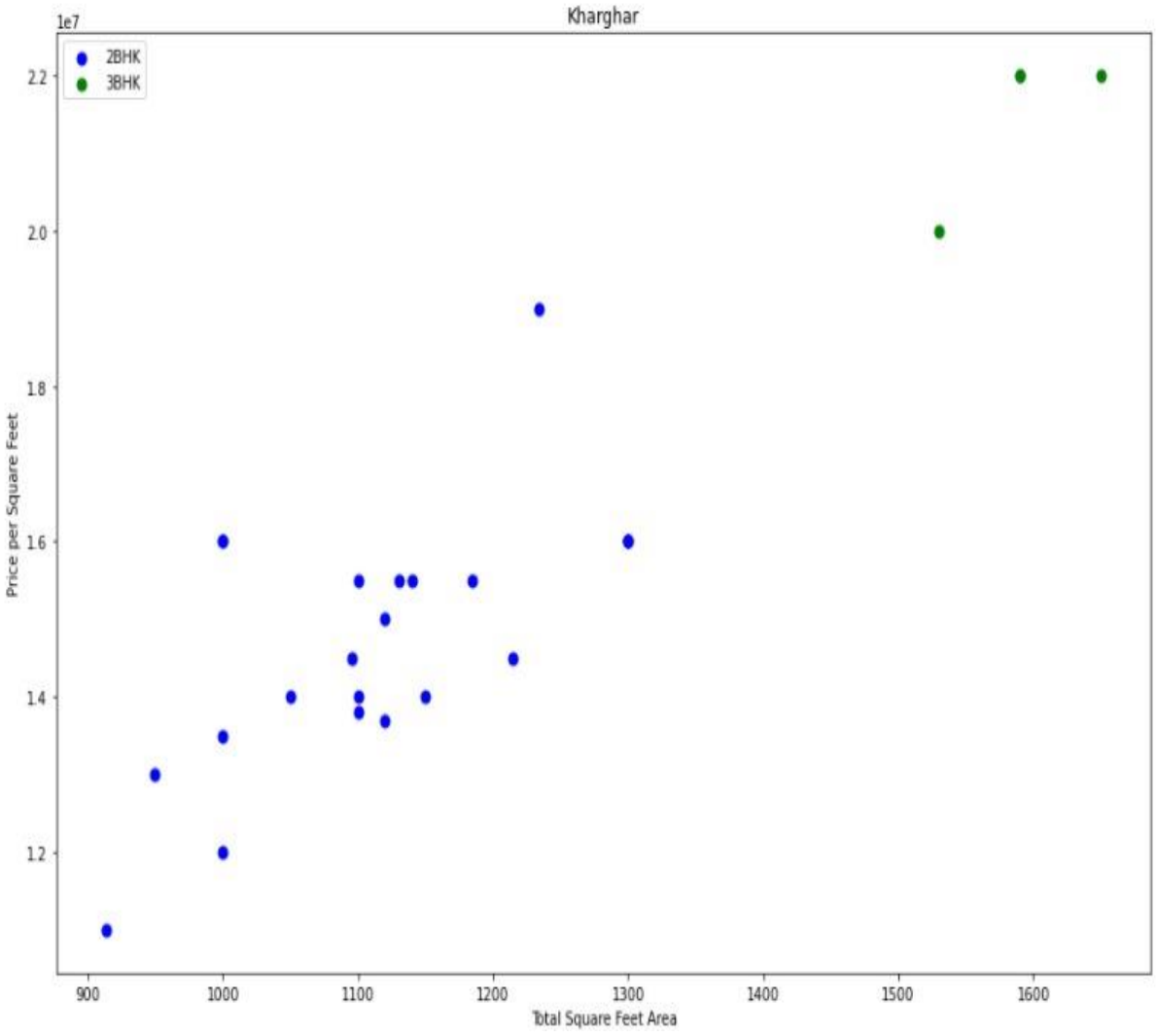
| | Price | Area | Location | size(rk or bhk) | Num_of_bedrooms | CarParking | Gasconnection | Children'splayarea |
|---|---|---|---|---|---|---|---|---|
| 0 | 4850000 | 720 | Kharghar | bhk | 1 | 1 | 0 | 0 |
| 1 | 4500000 | 600 | Kharghar | rk | 0 | 1 | 0 | 0 |
| 2 | 6700000 | 650 | Kharghar | rk | 0 | 1 | 0 | 1 |
| 3 | 4500000 | 650 | Kharghar | rk | 0 | 1 | 0 | 0 |
| 4 | 5000000 | 665 | Kharghar | rk | 0 | 1 | 0 | 0 |
| 5 | 17000000 | 2000 | Kharghar | bhk | 4 | 1 | 0 | 1 |
| 6 | 12500000 | 1550 | Kharghar | bhk | 3 | 1 | 1 | 0 |
| 7 | 10500000 | 1370 | Sector-13 Kharghar | bhk | 3 | 1 | 0 | 0 |

In [6]:
```
df3 = df2.dropna()
df3.isnull().sum()
```

Out[6]:
```
Price               0
Area                0
Location            0
size(rk or bhk)     0
Num_of_bedrooms     0
CarParking          0
Gasconnection       0
Children'splayarea  0
dtype: int64
```

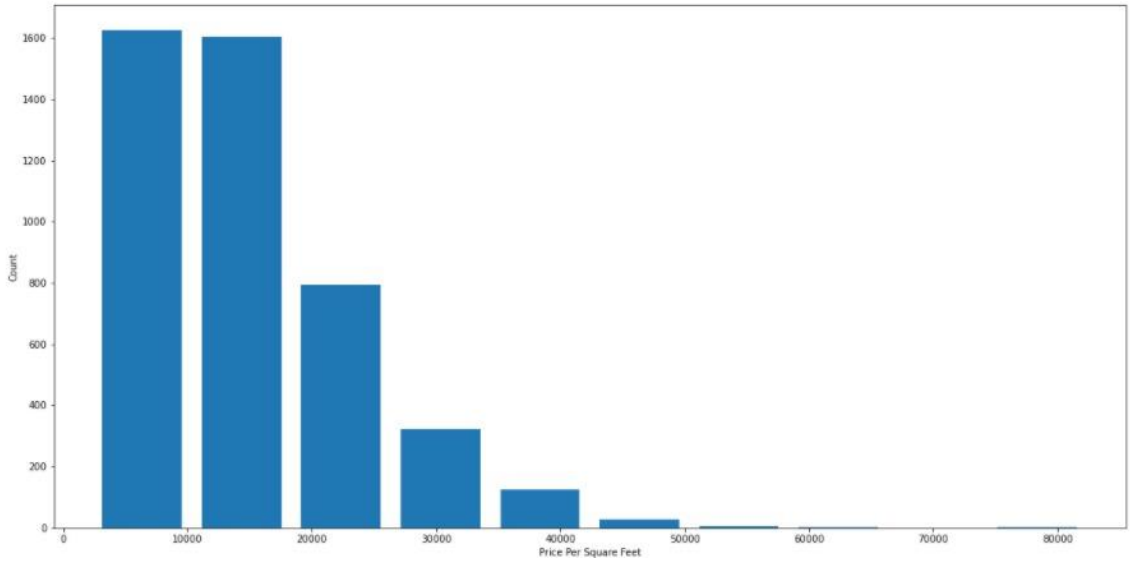
In [6]:
```
df3 = df2.dropna()
df3.isnull().sum()
```

Out[6]:
```
Price                0
Area                 0
Location             0
size(rk or bhk)      0
Num_of_bedrooms      0
CarParking           0
Gasconnection        0
Children'splayarea   0
dtype: int64
```



In [36]:
```
import matplotlib
matplotlib.rcParams["figure.figsize"]=(20,10)
plt.hist(df8.price_per_sqft,rwidth=0.8)
plt.xlabel("Price Per Square Feet")
plt.ylabel("Count")
```

Out[36]:
```
Text(0, 0.5, 'Count')
```



# Chapter 5

# Conclusion

This paper examined and analyzed the current research on the significant attributes of house price and analyzed the data mining techniques used to predict house price.

Technically, houses with a strategic location such as the accessibility to shopping mall or other facilities tend to be more expensive than houses in rural areas with limited numbers of facilities.

The accurate prediction model would allow investors or house buyers to determine the realistic price of a house as well as the house developers to decide the affordable house price. This paper addressed the attributes used by previous researchers to forecast a house price using various prediction models.

Taken together, the results of the survey have shown the potential of Machine Learning algorithms like XG-Boost in predicting house prices.

# Chapter 6

# Acknowledgment

It is a privilege for us to have been associated with Prof. Shashikant Patil, guide, during this project work. We have been greatly benefited by their valuable suggestions and ideas. It is with great pleasure that we express our deep sense of gratitude to them for their valuable guidance, constant encouragement and patience throughout this work.

We express our gratitude to Dr. B. R. Patil, Principal, Prof. Abhijeet Patil, Head of Department of Computer Engineering for their constant encouragement, co-operation, and support..

We take this opportunity to thank all our classmates for their company during the course work and for useful discussion we had with them.

We would be failing in our duties if we do not make a mention of our family members including our parents for providing moral support, without which this work would not have been completed.

Swapnil Patil

Ruturaj Raut

Avdhut Shinde

# Chapter 7

# References


1. Annina S, Mahima SD, Ramesh B. An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering (IJESE). 2015 January; I(1): 22-24.
2. David HW, William GM. No Free Lunch Theorems for Optimisation. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. 1997 April; I(1): 67-82.
3. Svensk Mäklarstatistik. [Online].; 2020. Available from: www.maklarstatistik.se..
4. Uyanık GK GN. A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences. 2013 Dec ; 106(1): 234-240.
5. Peter JB, Bo L. Regularization in Statistics. Sociedad de Estad´ıstica e Investigaci´on Operativa. 2006; XV(2): 271-344.
6. R T. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996 January; 58(1): 267-288.
   Fonti V. Feature Selection using LASSO. VU Amsterdam Research Paper in Business Analytics. 2017 Mars: p. 1-25.
7. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York: Springer series in statistics. 2001.
8. Clark AE, Troskie CG. Ridge Regression – A Simulation Study. Communications in Statistics - Simulation and. 2006: p. 605-619.
9. Yahya WB OJ. A note on ridge regression modeling. Electronic Journal of Applied Statistical Analysis. 2014 Oct : p. 343-361.
10. Ben Ishak A. Variable selection using support vector regression and random forests: A comparative study. Intelligent Data Analysis. 2016 January: p. 83-104.