VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018

A Mini Project Report on

"HOUSE PRICE PREDICTION SYSTEM"

Submitted in partial fulfillment of the requirements as a part of the

AI/ML INTERNSHIP

(NASTECH)

For the award of degree of

Bachelor of Engineering
in
Information Science and Engineering

Submitted by

VARUN DS 1RN18IS120

YUKTHA M 1RN18IS126

Internship Project Coordinators

Dr. R Rajkumar
Associate Professor
Dept. of ISE, RNSIT

Dr. S Satish Kumar
Professor
Dept. of ISE, RNSIT

Department of Information Science and Engineering

RNS Institute of Technology

Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post,
Bengaluru – 560 098

2021 -2022

RNS Institute of Technology

Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post,
Bengaluru – 560 098

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the mini project report entitled ***HOUSE PRICE PREDICTION SYSTEM*** has been successfully completed by **VARUN DS** bearing USN **1RN18IS120** and **Yuktha M** bearing USN **1RN18IS126** , presently VII semester students of **RNS Institute of Technology** in partial fulfillment of the requirements as a part of the ***AI/ML Internship (NASTECH)*** for the award of the degree of ***Bachelor of Engineering in Information Science and Engineering*** under **Visvesvaraya Technological University, Belagavi** during academic year **2021 – 2022**. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report and deposited in the departmental library. The mini project report has been approved as it satisfies the academic requirements as a part of Internship.

Dr. R Rajkumar	**Dr. S Satish Kumar**	**Dr. Suresh L**
Coordinator	Guide	Professor and HoD
Associate Professor	Professor	

External Viva

Name of the Examiners **Signature with date**

1. ______________________________ ____________________________

2. ______________________________ ____________________________

ABSTRACT

The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors. Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the industry

House price prediction project focuses on forecasting the coherent house prices for non-house holders based on their financial provisions and their aspirations. By analyzing the foregoing merchandise, fare ranges and also forewarns developments, speculated prices will be estimated. The motive of this paper is to help the seller to estimate the selling cost of a house perfectly and to help people to predict the exact time slap to accumulate a house. Some of the related factors that impact the cost were also taken into considerations such as physical conditions, concept and location etc.

House price prediction on a data set has been done by using linear regression technique. Moreover, this project can be considered as a further step towards more evidence-based decision making for the benefit of these stakeholders. The project focuses on assessment value for residential properties in Calgary between 2017-2020. The aim of our project is to build a predictive model for change in house prices in the year 2021 based on certain time and geography dependent variables.

ACKNOWLEDGMENT

At the very onset I would like to place our gratefulness to all those people who helped me in making the Internship a successful one.

Coming up, this internship to be a success was not easy. Apart from the sheer effort, the enlightenment of the very experienced teachers also plays a paramount role because it is they who guided me in the right direction.

First of all, I would like to thank the **Management of RNS Institute of Technology** for providing such a healthy environment for the successful completion of internship work.

In this regard, I express sincere gratitude to our beloved Principal **Dr. M K Venkatesha,** for providing us all the facilities.

We are extremely grateful to our own and beloved Professor and Head of Department of Information science and Engineering, **Dr. Suresh L**, for having accepted to patronize me in the right direction with all her wisdom.

We place our heartfelt thanks to **Dr. S Satish Kumar** Professor, Department of Information Science and Engineering for having guided internship and all the staff members of the department of Information Science and Engineering for helping at all times.

I thank **Mr. Deepak Garg, Founder, NASTECH**, for providing the opportunity to be a part of the Internship program and having guided me to complete the same successfully.

I also thank our internship coordinator **Dr. R Rajkumar,** Associate Professor, Department of Information Science and Engineering. I would thank my friends for having supported me with all their strength and might. Last but not the least, I thank my parents for supporting and encouraging me throughout. I have made an honest effort in this assignment.

VARUN D S	YUKTHA M
1RN18IS120	1RN18IS126

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 ORGANIZATION/INDUSTRY

1.1.1 COMPANY PROFILE

NASTECH is formed with the purpose of bridging the gap between Academia and Industry Nastech is one of the leading Global Certification and Training service providers for technical and management programs for educational institutions. We collaborate with educational institutes to understand their requirements and form a strategy in consultation with all stakeholders to fulfill those by skilling, reskilling and upskilling the students and faculties on new age skills and technologies.

1.1.2 DOMAIN/TECHNOLOGY

The domain chosen for our project is AI/ML. Machine learning, the fundamental driver of AI, is possible through algorithms that can learn themselves from data and identify patterns to make predictions and achieve your predefined goals, rather than blindly following detailed programmed instructions, like in traditional computer programming. This technology allows the machine to perceive, learn, reason and communicate through observation of data, like a child that grows up and acquires knowledge from examples. Machines also have the advantage of not being limited by our inherent biological limitations. With machine learning, manufacturing companies have increased production capacity up to 20%, while lowering material consumption rates by 4%.

Nowadays, the revolutionary AI technology evolved from rule-based expert systems to machine learning and more advanced subcomponents such as deep learning (learning representations instead of tasks), artificial neural networks (inspired by animal brains) and reinforcement learning (virtual agents rewarded if they made good decisions).

The AI can master the complexity of the intertwining industrial processes to enhance the whole flow of production instead of isolated processes. This enormous cognitive capacity gives the AI the ability to consider the spatial organization of plants and the timing constraints of live production. Another key advantage is the capability of AI algorithms to think probabilistically, with all the subtlety this allows in edge cases, instead of traditional rule-based methods that require rigid theories and a full comprehension of problems.

1.1.3 Department

R.N.Shetty Institute of Technology (RNSIT) established in the year 2001, is the brain-child of the Group Chairman, Dr. R. N. Shetty. The Murudeshwar Group of Companies headed by Sri. R. N. Shetty is a leading player in many industries viz construction, manufacturing, hotel, automobile, power & IT services and education. The group has contributed significantly to the field of education. A number of educational institutions are run by the

R. N. Shetty Trust, RNSIT being one amongst them. With a continuous desire to provide quality education to the society, the group has established RNSIT, an institution to nourish and produce the best of engineering talents in the country. RNSIT is one of the best and top accredited engineering colleges in Bengaluru.

1.2 PROBLEM STATEMENT

1.2.1 Existing System and their Limitations

A manual method is currently used in the market to predict the house price. The problem with this is that it doesn’t predict future prices of the houses mentioned by the customer. Due to this, the risk in investment in an apartment or an area increases considerably. To minimize this error, customers tend to hire an agent which again increases the cost of the process. Moreover, there is a chance that the agent might predict wrong estates and thus lead to loss of the customer’s investments. This leads to the modification and development of the existing system.

1.2.2 Proposed Solution

To eliminate the drawback of manual method, Machine learning algorithms can be used to help investors to invest in an appropriate estate according to their mentioned requirements. Also, the new system will be cost and time efficient. This will have simple operations. The proposed system works on Linear Regression Algorithm.

1.2.3 Program formulation

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

Chapter 2

REQUIREMENT ANALYSIS, TOOLS &TECHNOLOGIES

2.1 Hardware and Software Requirements

2.1.1 Hardware Requirements:

- Processor: Pentium IV or above
- RAM: GB or more
- Hard Disk: 2GB or more

2.1.2 Software Requirements:

- Operating System: Windows 7 or above
- IDE: Google Colab

2.2 Tools/Languages/Platforms

- Python

2.3 Literature Survey

1) **Predicting Housing Sales in Turkey Using Arima, LSTM And Hybrid Models** written by Ayşe Soy Temür, Melek Akgün, Günay Temür in the year 2019.
   In this study, a 124-month data set belonging to the 2008 (1) - 2018 (4) period has been taken into account for total housing sales in Turkey. In order to estimate the time series of sales, ARIMA (Auto Regressive Integrated Moving Average as linear model), LSTM (Long Short-Term Memory as nonlinear model) has been used. As to increase the estimation, a HYBRID (LSTM and ARIMA) model created has been used in the application. When MAPE (Mean Absolute Percentage Error*)* and MSE (Mean Squared Error) values obtained from each of these methods were compared, the best performance with the lowest error rate proved to be the HYBRID model, and the fact that all the application models have very close results shows the success of predictability.
2) **House Price Prediction Using Machine Learning and Neural Networks** written by Ayush Varma, Abhijit Sarma, Sagar Doshi and Rohini Nair in the year 2018. In this paper we aim to make our evaluations based on every basic parameter that is considered while determining the price. We use various regression techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most

accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied.

3) **House Price Prediction Modelling Using Machine Learning** written by Dr. M. Thamarai and Dr. S P. Malarvizhi in the year 2020. Proposed work makes use of the attributes or features of the houses such as number of bedrooms available in the house, age of the house, travelling facility from the location, school facility available nearby the houses and shopping malls available nearby the house location. House availability based on desired features of the house and house price prediction are modeled in the proposed work and the model is constructed for a small town in West Godavari district of Andhrapradesh. The work involves decision tree classification, decision tree regression and multiple linear regression and is implemented using Scikit-Learn Machine Learning Tool.

4) **House Price Prediction Using Machine Learning Algorithms** written by Naalla Vineeth, Maturi Ayyappa and B. Bharathi in the year 2018. Due to increase in urbanization, there is an increase in demand for renting houses and purchasing houses. Therefore, to determine a more effective way to calculate house price that accurately reflects the market price becomes a hot topic. The paper focuses on finding the house price accurately by using machine learning algorithms like simple linear regression (SLR), Multiple linear regression (MLR), Neural Networks (NN). The algorithm which has the lower Mean Square Error (MSE) is chosen as the best algorithm for predicting the house price. This will be helpful for both the sellers and buyers for finding the best price for the house.

5) **House Price Prediction Using Regression Techniques: A Comparative Study** written by CH. Raga Madhuri, G. Anuradha and M. Vani Pujitha in the year 2019. The objective of the paper is to forecast the coherent house prices for non-house holders based on their financial provisions and their aspirations. By analyzing the foregoing merchandise, fare ranges and also forewarns developments, speculated prices will be estimated. The paper involves predictions using different Regression techniques like Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting and Ada Boost Regression. House price prediction on a data set has been done by using all the above-mentioned techniques to find out the best among them. The motive of this paper is to help the seller to estimate the selling cost of a house perfectly and to help people to predict the exact time slap to accumulate a house.

6) **Prediction of House Price Based on The Back Propagation Neural Network in The Keras Deep Learning Framework** written by Zhongyun Jiang and Guoxin Shen in the year 2019. This paper uses the housing data of the chain home network to predict the price of second-hand housing in Shanghai. Firstly, this paper use the crawler technology to parse the URL text information through the j son request address and the BeautifulSoup parser. Then a multi-layer feedforward neural network model trained by error inverse propagation algorithm is established based on the deep learning library Keras. Finally, to enter standardized data to predict the price. The experimental results show that for the model with Gaussian noise, the sample with an absolute value of the relative error between the predicted value and the actual value is 95.59%.

7) **House Prices Prediction with Machine Learning Algorithms** written by Chenchen Fan, Zechen Cui and Xiaofeng Zhong in the year 2018. Based on the data set compiled by D. D. Cock and the competition run by kaggle.com, we propose a house prices prediction algorithm in Ames, lowa by deliberating on data processing, feature engineering and combination forecasting. Our prediction ranks the 35th of the total 2221 results on the public leaderboard of Kaggle.com and the RMSE of predicted results after taking logarithm from all the test data is 0.12019, which shows good performance and small of over-fitting.

Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia written by Danh Phan in the year 2018. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in Australia 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 Melbourne. Moreover, experiments demonstrate that the combination of Stepwise and Support Vector Machine that is based on mean squared error measurement is a competitive approach

Chapter 3

DESIGN AND IMPLIMENTATION

3.1 Architecture/ DFD/Sequence diagram/Class diagrams /Flowchart

Keras Neural network has been used in the project which is a fast, open-source, and easy-to-use Neural Network Library written in Python.

Since there are 19 features, 19 neurons are inserted as a start, 4 hidden layers and 1 output layer due to predict house Price.

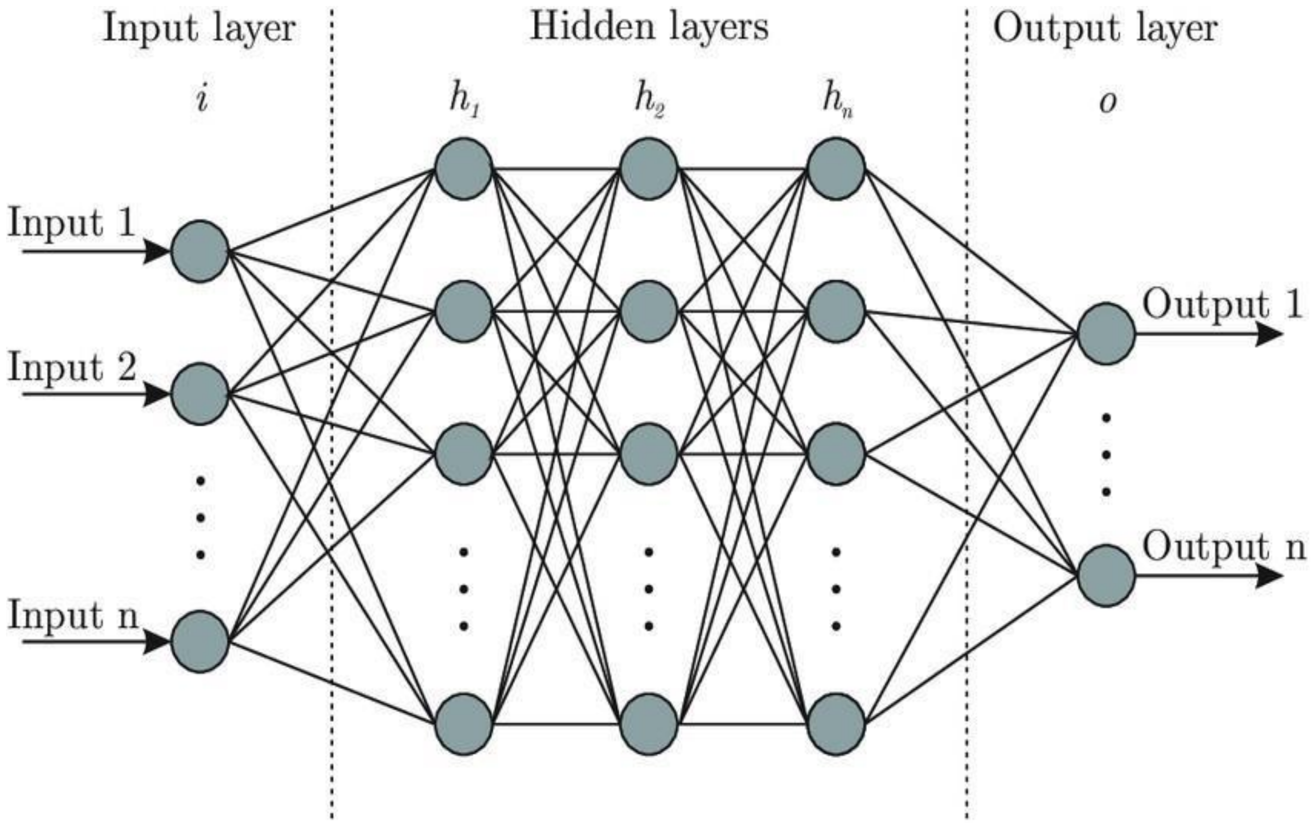


Figure 3.1 Typical Keras Neural Model

In the above fig3.1, we discuss about Keras Neural Model.

Steps to perform Keras Regression:

Step 1

First download and import the dataset using pandas

Step 2

Clean the data:The dataset contains a few unknown values

Step 3

Split the dataset into a training set and a test set. You will use the test set in the final evaluation of your models.

Step 4

Review the joint distribution of a few pairs of columns from the training set.

Step 5

Separate the target value—the "label"—from the features. This label is the value that you will train the model to predict.

Step 6

Normalize features that use different scales and ranges.

KERAS

Keras is an API used for running high-level neural networks. The model runs on top of TensorFlow, and was developed by Google.

The main competitor to Keras at this point in time is PyTorch, developed by Facebook. While PyTorch has a somewhat higher level of community support, it is a particularly verbose language and I personally prefer Keras for greater simplicity and ease of use in building and deploying models.

NEURAL NETWORK

Neural networks (NN), also called artificial neural networks (ANN) are a subset of learning algorithms within the machine learning field that are loosely based on the concept of biological neural networks.A neural network is a computational system that creates predictions based on existing data. Let us train and test a neural network using the neuralnet library in R.

A neural network consists of:

Input layers: Layers that take inputs based on existing data

Hidden layers: Layers that use backpropagation to optimise the weights of the input variables in order to improve the predictive power of the model

Output layers: Output of predictions based on the data from the input and hidden layers

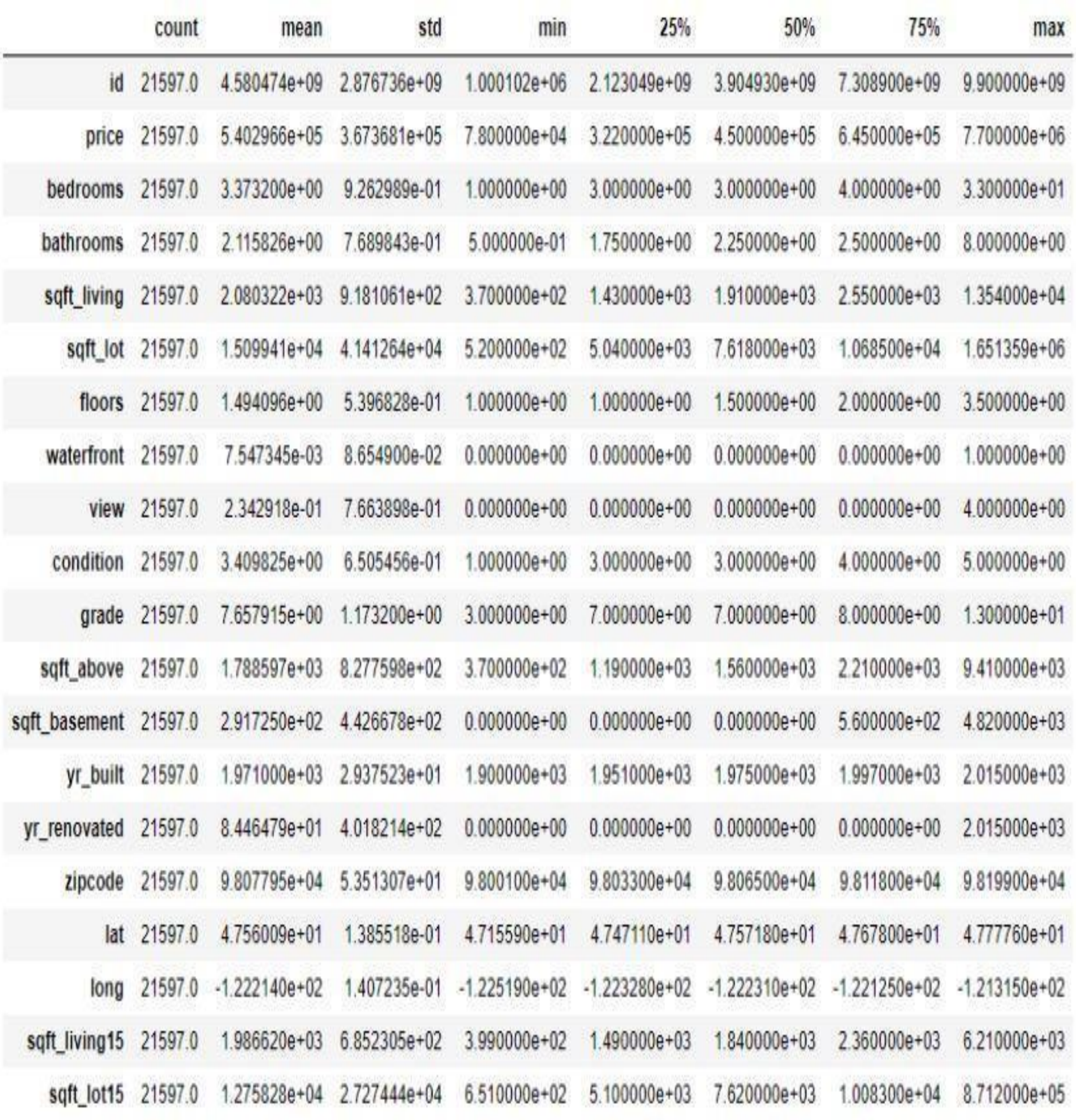
3.2 Solution

The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house price. Objective is to predict the house price.

Keras model has been used in terms of minimizing the difference between predicted and actual rating.

The following features have been used:

1. Date: Date house was sold
2. Price: Price is prediction target
3. Bedrooms: Number of Bedrooms/House
4. Bathrooms: Number of bathrooms/House
5. Sqft_Living: square footage of the home
6. Sqft_Lot: square footage of the lot
7. Floors: Total floors (levels) in house
8. Waterfront: House which has a view to a waterfront
9. View: Has been viewed
10. Condition: How good the condition is (Overall)
11. Grade: grade given to the housing unit, based on King County grading system
12. Sqft_Above: square footage of house apart from basement
13. Sqft_Basement: square footage of the basement
14. Yr_Built: Built Year
15. Yr_Renovated: Year when house was renovated
16. Zipcode: Zip
17. Lat: Latitude coordinate
18. Long: Longitude coordinate
19. Sqft_Living15: Living room area in 2015(implies — some renovations)
20. Sqft_Lot15: lotSize area in 2015(implies — some renovations)



	count	mean	std	min	25%	50%	75%	max
id	21597.0	4.580474e+09	2.876736e+09	1.000102e+06	2.123049e+09	3.904930e+09	7.308900e+09	9.900000e+09
price	21597.0	5.402966e+05	3.673681e+05	7.800000e+04	3.220000e+05	4.500000e+05	6.450000e+05	7.700000e+06
bedrooms	21597.0	3.373200e+00	9.262989e-01	1.000000e+00	3.000000e+00	3.000000e+00	4.000000e+00	3.300000e+01
bathrooms	21597.0	2.115826e+00	7.689843e-01	5.000000e-01	1.750000e+00	2.250000e+00	2.500000e+00	8.000000e+00
sqft_living	21597.0	2.080322e+03	9.181061e+02	3.700000e+02	1.430000e+03	1.910000e+03	2.550000e+03	1.354000e+04
sqft_lot	21597.0	1.509941e+04	4.141264e+04	5.200000e+02	5.040000e+03	7.618000e+03	1.068500e+04	1.651359e+06
floors	21597.0	1.494096e+00	5.396828e-01	1.000000e+00	1.000000e+00	1.500000e+00	2.000000e+00	3.500000e+00
waterfront	21597.0	7.547345e-03	8.654900e-02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
view	21597.0	2.342918e-01	7.663898e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	4.000000e+00
condition	21597.0	3.409825e+00	6.505456e-01	1.000000e+00	3.000000e+00	3.000000e+00	4.000000e+00	5.000000e+00
grade	21597.0	7.657915e+00	1.173200e+00	3.000000e+00	7.000000e+00	7.000000e+00	8.000000e+00	1.300000e+01
sqft_above	21597.0	1.788597e+03	8.277598e+02	3.700000e+02	1.190000e+03	1.560000e+03	2.210000e+03	9.410000e+03
sqft_basement	21597.0	2.917250e+02	4.426678e+02	0.000000e+00	0.000000e+00	0.000000e+00	5.600000e+02	4.820000e+03
yr_built	21597.0	1.971000e+03	2.937523e+01	1.900000e+03	1.951000e+03	1.975000e+03	1.997000e+03	2.015000e+03
yr_renovated	21597.0	8.446479e+01	4.018214e+02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.015000e+03
zipcode	21597.0	9.807795e+04	5.351307e+01	9.800100e+04	9.803300e+04	9.806500e+04	9.811800e+04	9.819900e+04
lat	21597.0	4.756009e+01	1.385518e-01	4.715590e+01	4.747110e+01	4.757180e+01	4.767800e+01	4.777760e+01
long	21597.0	-1.222140e+02	1.407235e-01	-1.225190e+02	-1.223280e+02	-1.222310e+02	-1.221250e+02	-1.213150e+02
sqft_living15	21597.0	1.986620e+03	6.852305e+02	3.990000e+02	1.490000e+03	1.840000e+03	2.360000e+03	6.210000e+03
sqft_lot15	21597.0	1.275828e+04	2.727444e+04	6.510000e+02	5.100000e+03	7.620000e+03	1.008300e+04	8.712000e+05

Figure 3.2 Description of the Dataset

The above fig3.2, shows the description of the dataset.

`3.3 Algorithm

The independent values are taken along the x-axis and dependent along the y-axis.

```
1. Read n //total number of points
2. Read x, y //x and y co-ordinates of points
3. Initialize diffx[n], diffy[n]
4. Initialize diffxy, diffx2 to 0
5. for i = 1 to n do
     calculate the mean of x: xm mean of y: ym
     diffx[i] = x[i] – xm //find the difference values between each x and mean of x
     diffy[i] = y[i] – ym //find the difference values between each y and mean of y
         diffx2 = Σ(diffx[i])2 //calculate the summation of all the difference values of x
         diffxy = Σ((diffx[i]) * (diffy[i])) //compute the product of diff values of x and y
   end for
6. m = diffxy / diffx2 //the slope value is obtained by this Formula
7. c = ym – (m * xm) //the intercept value is obtained with this Formula
8. Equation complete: y = (m * x) + c
9. Stop.
```

By substituting the value of x in the obtained equation the respective y value can be found

3.4 Libraries

- Pandas
- Numpy
- Seaborn
- Matplotlib

Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis/manipulation tool available in any language**.** It is already well on its way toward this goal.

Numpy

NumPy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python.

Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.
- Customize visual style and layout.
- Export to many file formats .
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

Seaborn

Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

Chapter 4

OBSERVATION AND RESULTS

4.1 Testing

Evaluation on Test Data

```
y_pred = model.predict(X_test)
from sklearn import metric
sprint ('MAE:', metrics.mean_absolute_error(y_test, y_pred))
print('MSE:', metrics.mean_squared_error(y_test, y_pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test,y_pred)))
print('VarScore:',metrics.explained_variance_score(y_test,y_pred))
```

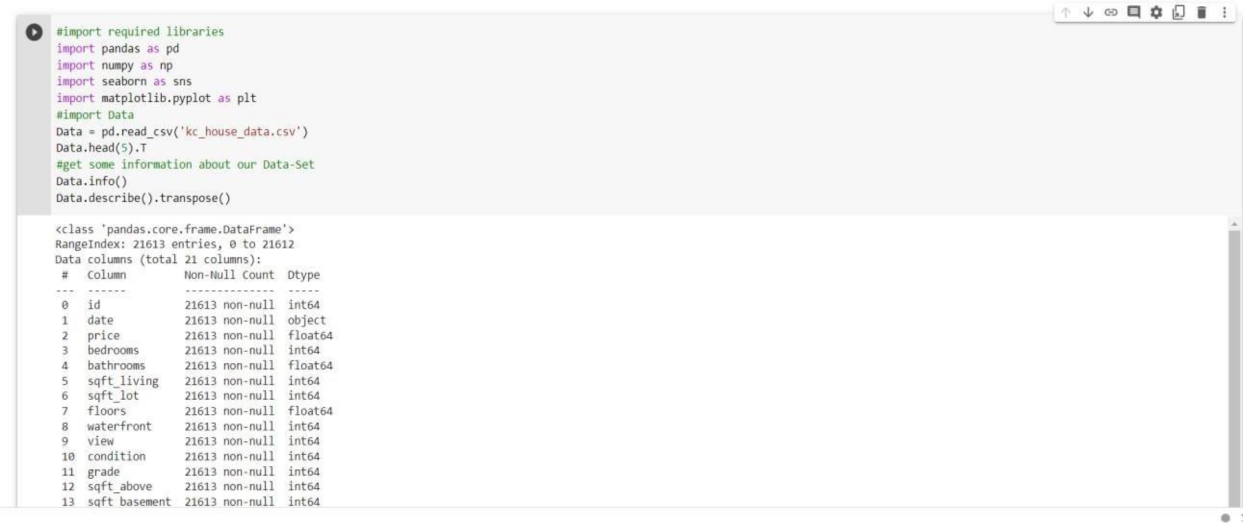
Visualizing Our predictions

```
fig = plt.figure(figsize=(10,5))
plt.scatter(y_test,y_pred)
```

Perfect predictions

```
plt.plot(y_test,y_test,'r')
```

4.2 Results & Snapshots



```
#import required libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
#import Data
Data = pd.read_csv('kc_house_data.csv')
Data.head(5).T
#get some information about our Data-Set
Data.info()
Data.describe().transpose()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             21613 non-null  int64
 1   date           21613 non-null  object
 2   price          21613 non-null  float64
 3   bedrooms       21613 non-null  int64
 4   bathrooms      21613 non-null  float64
 5   sqft_living    21613 non-null  int64
 6   sqft_lot       21613 non-null  int64
 7   floors         21613 non-null  float64
 8   waterfront     21613 non-null  int64
 9   view           21613 non-null  int64
 10  condition      21613 non-null  int64
 11  grade          21613 non-null  int64
 12  sqft_above     21613 non-null  int64
 13  sqft_basement  21613 non-null  int64
```

Figure 4.1 Reading CSV File

In the above fig 4.1, we are first importing all the modules required and then reading the dataset.csv file.

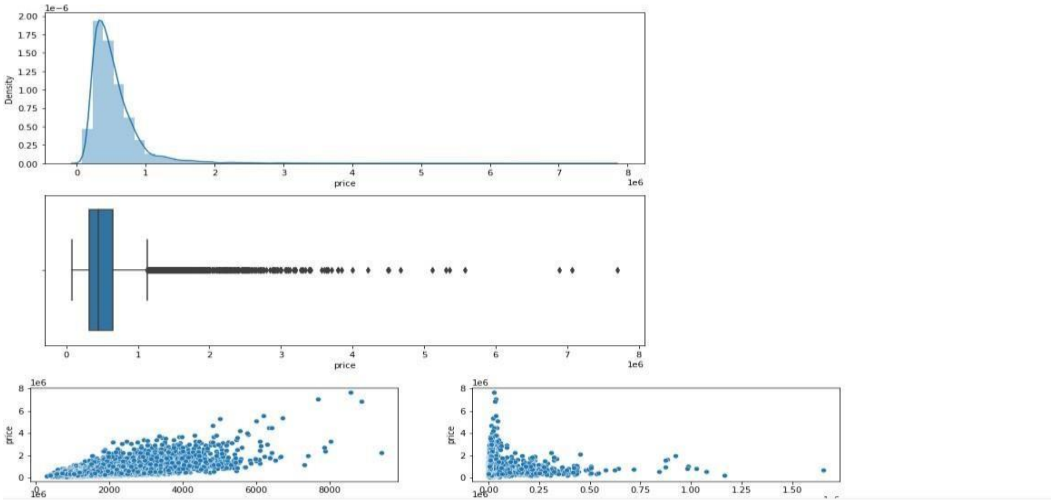


Figure 4.2 Visualizing house price

In the above fig4.2, we are visualizing the house prices from the given dataset in distplot and boxplot so that it will be easy to understand the price range. With distribution plot of price, we can visualize that most of the prices are between 0 and around 1M with few outliers close to 8 million. It would make sense to drop those outliers in our analysis.

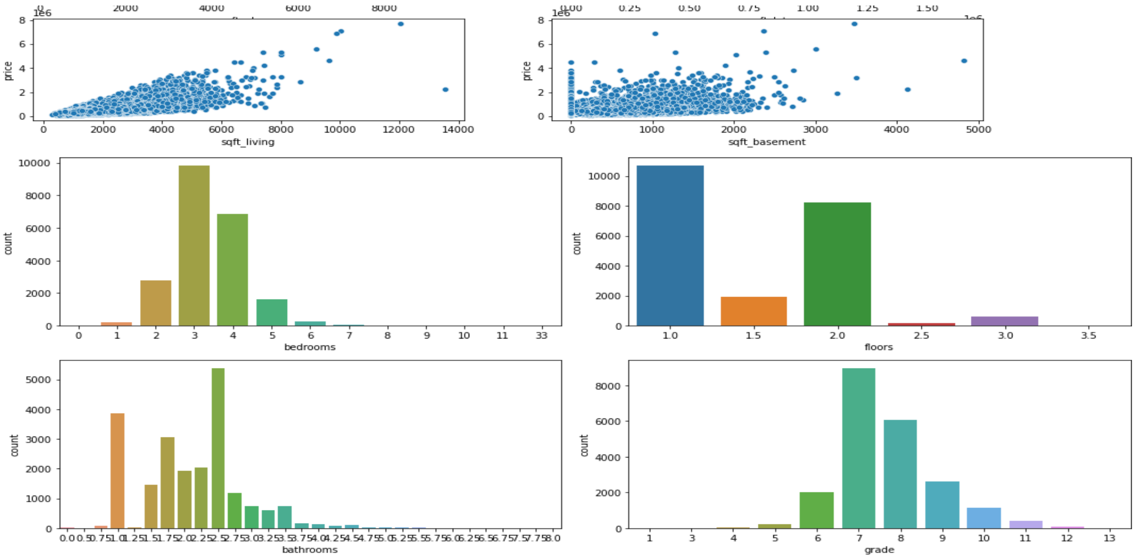
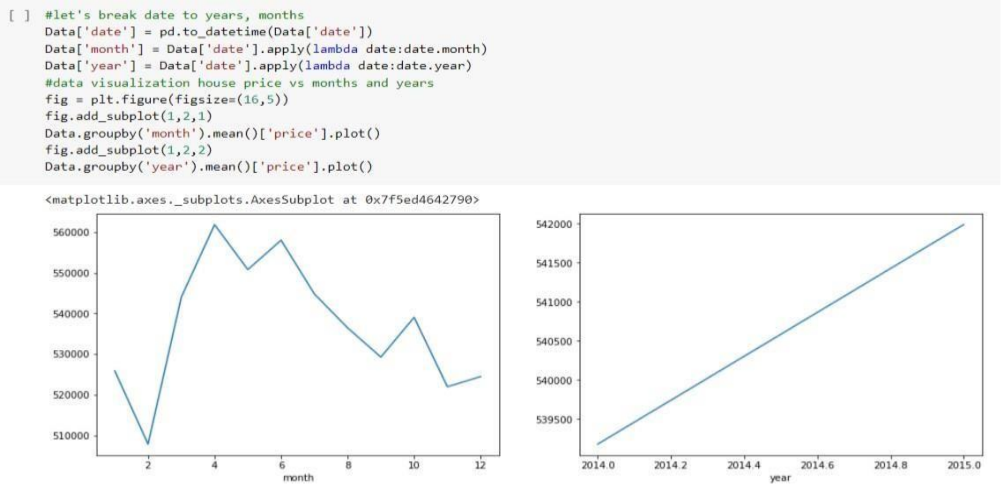


Figure 4.3 Count of bathroom, floor and bedrooms

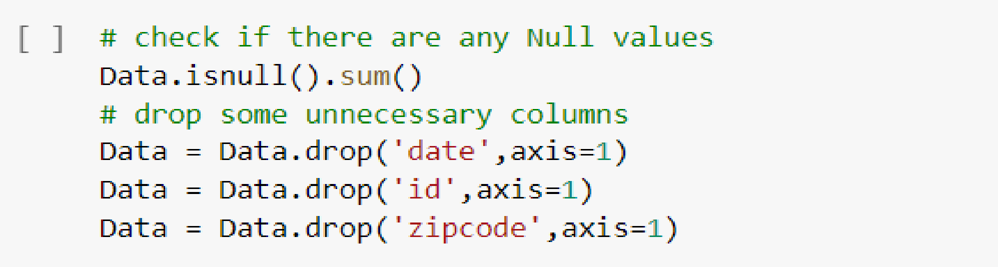
In the above fig4.3, we are visualizing the count of bedrooms, floors and bathrooms from the given dataset. We can see the most common and least common type of amenities in the house.



```
[ ]  #let's break date to years, months
     Data['date'] = pd.to_datetime(Data['date'])
     Data['month'] = Data['date'].apply(lambda date:date.month)
     Data['year'] = Data['date'].apply(lambda date:date.year)
     #data visualization house price vs months and years
     fig = plt.figure(figsize=(16,5))
     fig.add_subplot(1,2,1)
     Data.groupby('month').mean()['price'].plot()
     fig.add_subplot(1,2,2)
     Data.groupby('year').mean()['price'].plot()

     <matplotlib.axes._subplots.AxesSubplot at 0x7f5ed4642790>
```

Figure 4.4 Price range month vs year

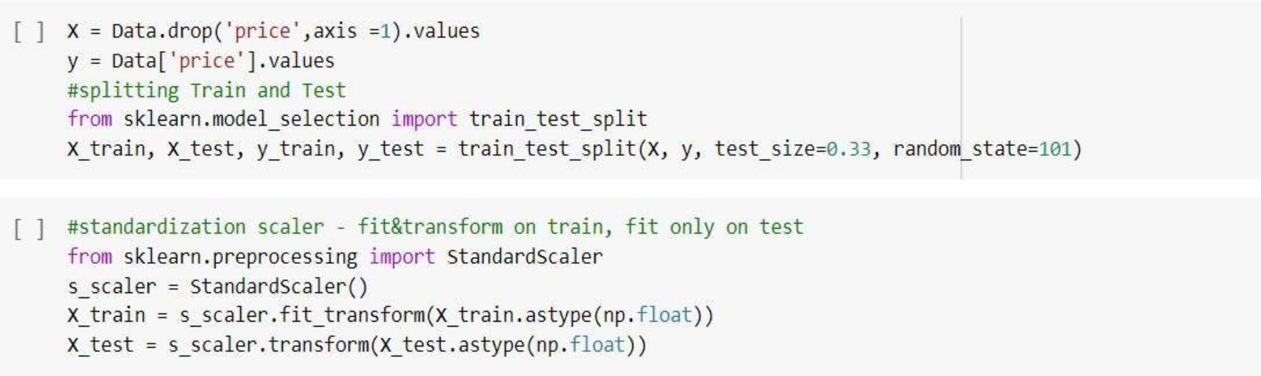
In the above fig4.4, we are viewing the price change that is happening monthly and price change that is happening yearly.



```
[ ]  # check if there are any Null values
     Data.isnull().sum()
     # drop some unnecessary columns
     Data = Data.drop('date',axis=1)
     Data = Data.drop('id',axis=1)
     Data = Data.drop('zipcode',axis=1)
```

Figure 4.5 Removing unnecessary columns

In the above fig4.5, we are first checking if there are any null values and then removing all the unnecessary columns from the dataset.



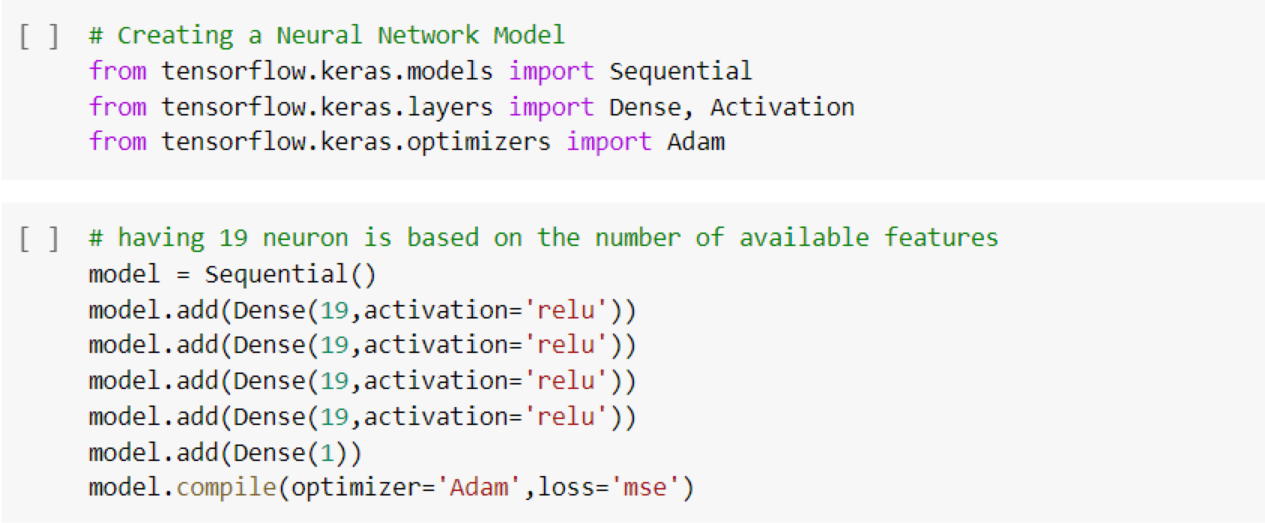
```
[ ]  X = Data.drop('price',axis =1).values
     y = Data['price'].values
     #splitting Train and Test
     from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=101)

[ ]  #standardization scaler - fit&transform on train, fit only on test
     from sklearn.preprocessing import StandardScaler
     s_scaler = StandardScaler()
     X_train = s_scaler.fit_transform(X_train.astype(np.float))
     X_test = s_scaler.transform(X_test.astype(np.float))
```

Figure 4.6 Dataset Preparation

Features(X): The columns that are inserted into our model will be used to make predictions.

Prediction (y): Target variable that will be predicted by the features.

Feature scaling will help us see all the variables from the same lens (same scale), it will also help our models learn faster.



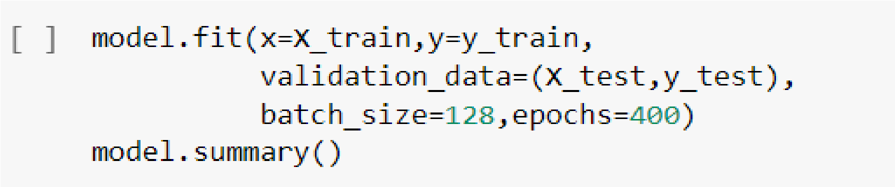
```
[ ]  # Creating a Neural Network Model
     from tensorflow.keras.models import Sequential
     from tensorflow.keras.layers import Dense, Activation
     from tensorflow.keras.optimizers import Adam
```

```
[ ]  # having 19 neuron is based on the number of available features
     model = Sequential()
     model.add(Dense(19,activation='relu'))
     model.add(Dense(19,activation='relu'))
     model.add(Dense(19,activation='relu'))
     model.add(Dense(19,activation='relu'))
     model.add(Dense(1))
     model.compile(optimizer='Adam',loss='mse')
```

Figure 4.7 Keras Implementation

In the above fig4.7, we first import all the keras models. Since we have 19 features, let's insert 19 neurons as a start, 4 hidden layers and 1 output layer due to predict house Price.

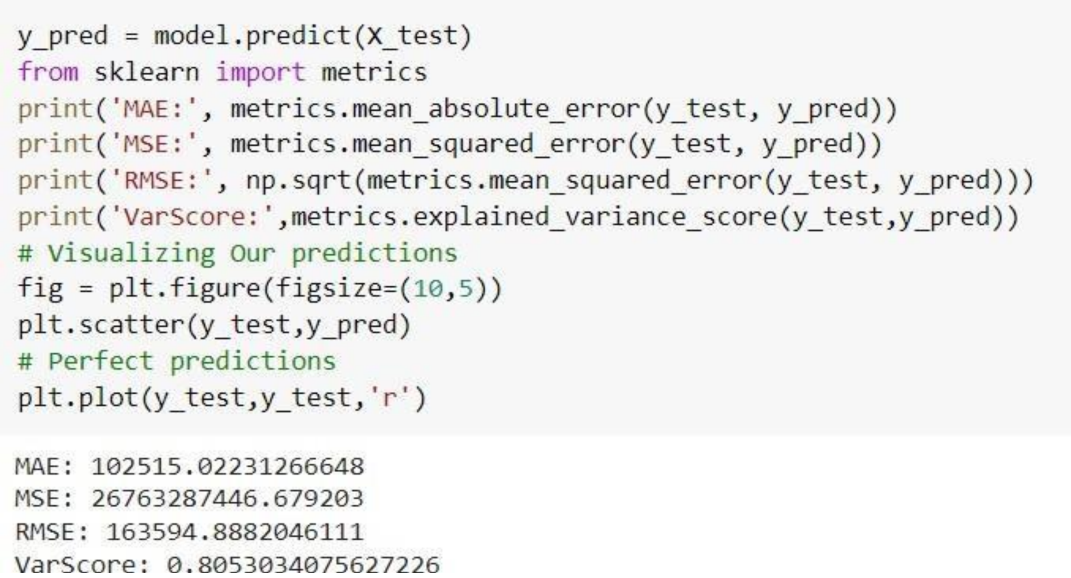
Also, ADAM optimization algorithm is used for optimizing loss function (Mean squared error).



```
[ ]  model.fit(x=X_train,y=y_train,
               validation_data=(X_test,y_test),
               batch_size=128,epochs=400)
     model.summary()
```

Figure 4.8 Validating Accuracy

In the above fig4.8, we train the model for 400 epochs, and each time record the training and validation accuracy in the history object. To keep track of how well the model is performing for each epoch, the model will run in both train and test data along with calculating the loss function.



```
y_pred = model.predict(X_test)
from sklearn import metrics
print('MAE:', metrics.mean_absolute_error(y_test, y_pred))
print('MSE:', metrics.mean_squared_error(y_test, y_pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
print('VarScore:',metrics.explained_variance_score(y_test,y_pred))
# Visualizing Our predictions
fig = plt.figure(figsize=(10,5))
plt.scatter(y_test,y_pred)
# Perfect predictions
plt.plot(y_test,y_test,'r')
```

```
MAE: 102515.02231266648
MSE: 26763287446.679203
RMSE: 163594.8882046111
VarScore: 0.8053034075627226
```

Figure 4.9 Evaluation of Test Data

In the above fig4.9, we are calculating mean absolute error, mean square error and variable score to check the accuracy of this algorithm. It is clearly seen that this approach is 81% accurate.

Chapter 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1 Conclusion

Aim of the project is to predict the house price taking into consideration various features pertaining to a house such as number of bedrooms, bathrooms, floors, sqrt_area, waterfront, view, condition, grade etc. which has successfully been achieved with an accuracy of 81%. The proposed model is definitely the best substitute for the manual method wherein third party is involved and is potentially vulnerable along with being not so pocket friendly. Based on the results, it can be concluded that such ML-driven predictions are easily comprehendible and significant from a data-analytics point of view. When correctly implemented, a high rate of accuracy can be achieved.

5.2 Future Enhancement

To make the system even more informative and user-friendly, Gmap can be included. This will show the neighborhood amenities such as hospitals, schools surrounding a region of 1 km from the given location. This can also be included in making predictions since the presence of such factors increases the valuation of real estate property.

Various other machine learning algorithms can also be used apart from Keras to improve the accuracy of the model.

Chapter 6

REFERENCE

[1] A. S. Temür, M. Akgün, and G. Temür, “Predicting Housing Sales in Turkey Using Arima, Lstm and Hybrid Models,” J. Bus. Econ. Manag., vol. 20, no. 5, pp. 920–938, 2019, doi: 10.3846/jbem.2019.10190.

[2] Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing ICMLC 2018. doi:10.1145/3195106.3195133.

[3] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.

[4] Thamarai, M. & Malarvizhi, S. (2020). House Price Prediction Modeling Using Machine Learning. International Journal of Information Engineering and Electronic Business. 12. 15-20. 10.5815/ijieeb.2020.02.03.

[5] House Price Prediction Using Machine Learning Algorithms Soft Computing Systems, 2018, Volume 837 ISBN : 978-981-13-1935-8 Naalla Vineeth, Maturi Ayyappa, B. Bharathi

[6] C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.

[7] Z. Jiang and G. Shen, "Prediction of House Price Based on The Back Propagation Neural Network in The Keras Deep Learning Framework," 2019 6th International Conference on Systems and Informatics (ICSAI), 2019, pp. 1408-1412, doi:10.1109/ICSAI48974.2019.9010071.

[8] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.