

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

Mr. Deepak Garg

Founder

NASTECH



Department of Information Science and Engineering

RNS Institute of Technology

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

2020 -2021

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 Mobile.

Dr. R Rajkumar

Coordinator

Associate Professor

Mr. Deepak Garg

Coordinator

Dr. Suresh L

Professor and HoD

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

The fulfillment and rapture that go with the fruitful finishing of any assignment would be inadequate without the specifying the people who made it conceivable, whose steady direction and support delegated the endeavors with success.

We would like to profoundly thank **Management of RNS Institute of Technology** for providing such a healthy environment to carry out this AI/ML Internship Project.

We would like to express our thanks to our Principal **Dr. M K Venkatesha** for his support and inspired us towards the attainment of knowledge.

We wish to place on record our words of gratitude to **Dr. Suresh L**, Professor and Head of the Department, Information Science and Engineering, for being the enzyme and master mind behind our Mobile Application Development Laboratory with Mini Project Work.

We would like to express our profound and cordial gratitude to my Internship Project Coordinators, **Dr. R Rajkumar**, Associate Professor, Department of Information Science and Engineering for their valuable guidance, constructive comments, continuous encouragement throughout the Mini Project Work and guidance in preparing report.

We would like to thank all other teaching and non-teaching staff of Information Science & Engineering who have directly or indirectly helped us to carry out the Mini Project Work.

Also, we would like to acknowledge and thank our parents who are source of inspiration and instrumental in carrying out this Mini Project Work.

VARUN DS
USN:1RN18IS120

YUKTHA M
USN:1RN18IS126

TABLE OF CONTENTS

Abstract	iii
Acknowledgement	iv
Table of Content	v
List of Figures	vi
1. INTRODUCTION	01
1.1. ORGANIZATION/ INDUSTRY	01
1.1.1. Company Profile	01
1.1.2. Domain/ Technology (Data Science/Mobile computing/...)	01
1.1.3. Department/ Division / Group	02
1.2. PROBLEM STATEMENT	02
1.2.1. Existing System and their Limitations	02
1.2.2. Proposed Solution	02
1.2.3. Problem formulation	02
2. REQUIREMENT ANALYSIS, TOOLS & TECHNOLOGIES	03
2.1. Hardware & Software Requirements	03
2.2. Tools/ Languages/ Platform	03
3. DESIGN AND IMPLIMENTATION	03
3.1. Architecture/ DFD/Sequence diagram/Class diagrams /Flowchart	03
3.2. Problem statement	04
3.3. Algorithm/Methods/ Pseudo code	06
3.4. Libraries used / API'S	06
4. OBSERVATIONS AND RESULTS	07
4.1. Testing	07
4.2. Results & Snapshots	07
5. CONCLUSION AND FUTURE WORK	11
5.1. Conclusion	11
5.2. Future Enhancement	11
6. REFERENCES	12

LIST OF FIGURES

Figure. No.	Descriptions	Page
Figure. 3.1	Typical Keras Neural Model	04
Figure. 3.2	Description of the Dataset	05
Figure. 4.1	Reading CSV file	07
Figure. 4.2	Visualizing house price	08
Figure. 4.3	Count of bathroom, floor and bedrooms	08
Figure. 4.4	Price range month vs year	09
Figure. 4.5	Removing unnecessary columns	09
Figure. 4.6	Dataset preparation	09
Figure. 4.7	Keras Implementation	10
Figure. 4.8	Validating Accuracy	10
Figure. 4.9	Evaluation of test data	10

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

Chapter 3

DESIGN AND IMPLEMENTATION

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.

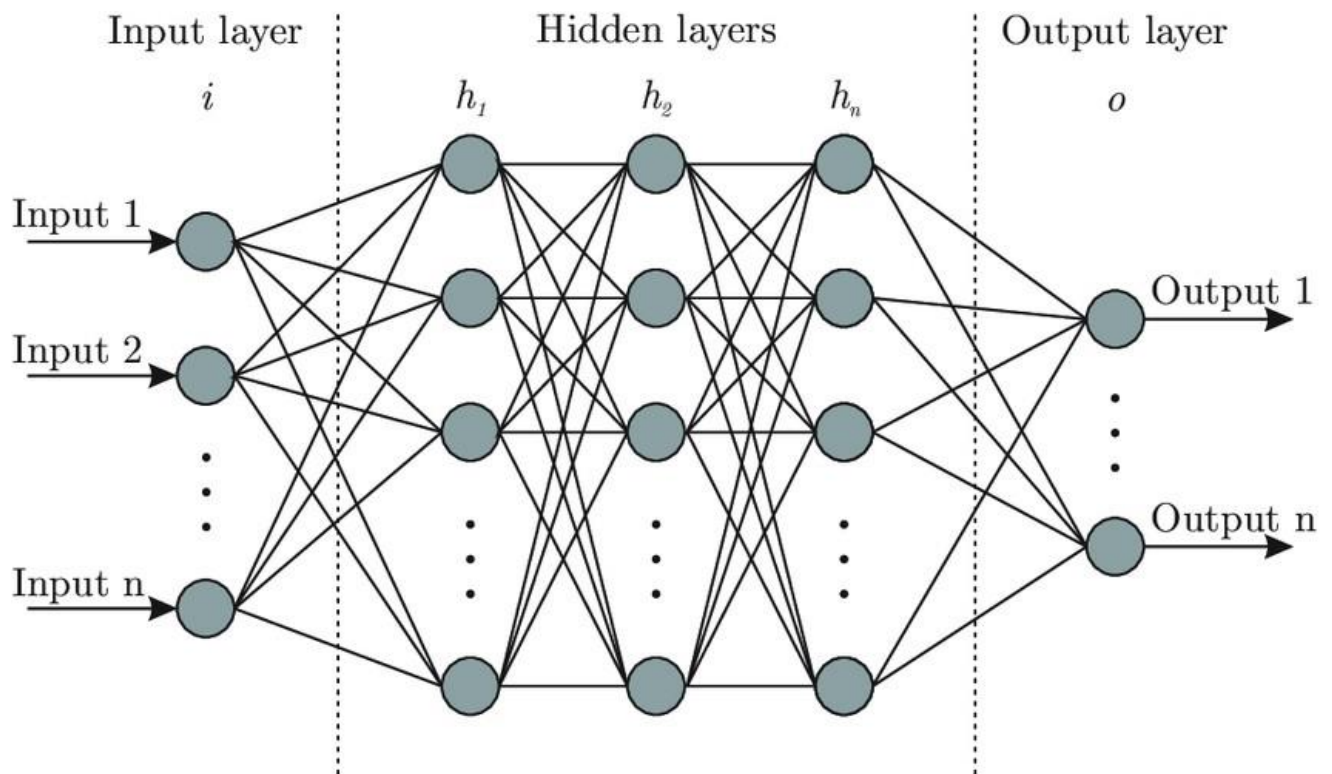


Figure 3.1 Typical Keras Neural Model

3.2 Problem Statement

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 = $\Sigma(\text{diffx}[i]^2)$ //calculate the summation of all the difference values of x
 diffxy = $\Sigma((\text{diffx}[i]) * (\text{diffy}[i]))$ //compute the product of diff values of x and y
end for
6. $m = \text{diffxy} / \text{diffx2}$ //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

Chapter 4

OBSERVATION AND RESULTS

4.1 Testing

Evaluation on Test Data

```
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))
```

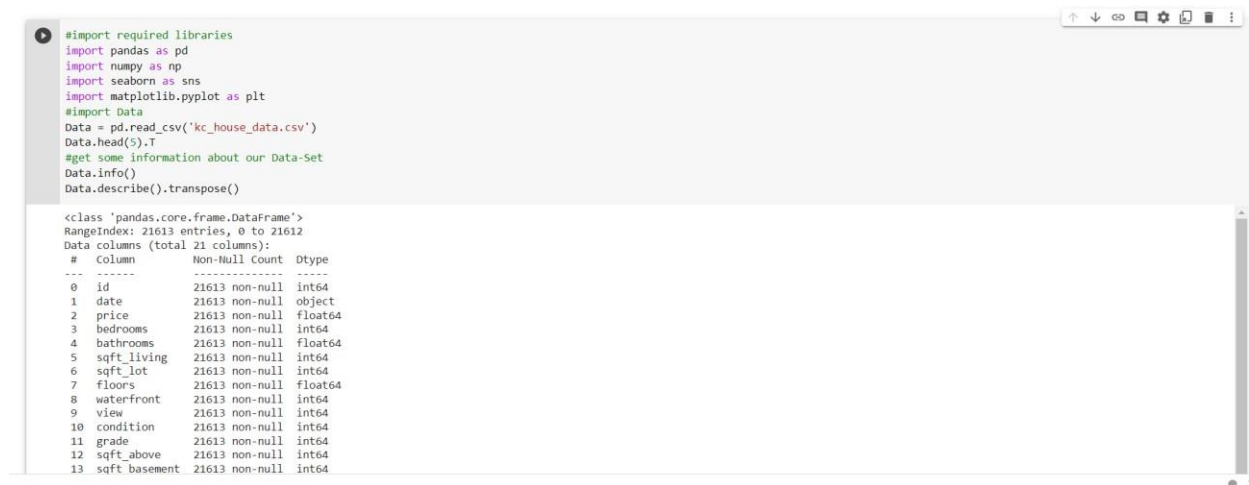
Visualizing Our predictions

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

Perfect predictions

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

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.

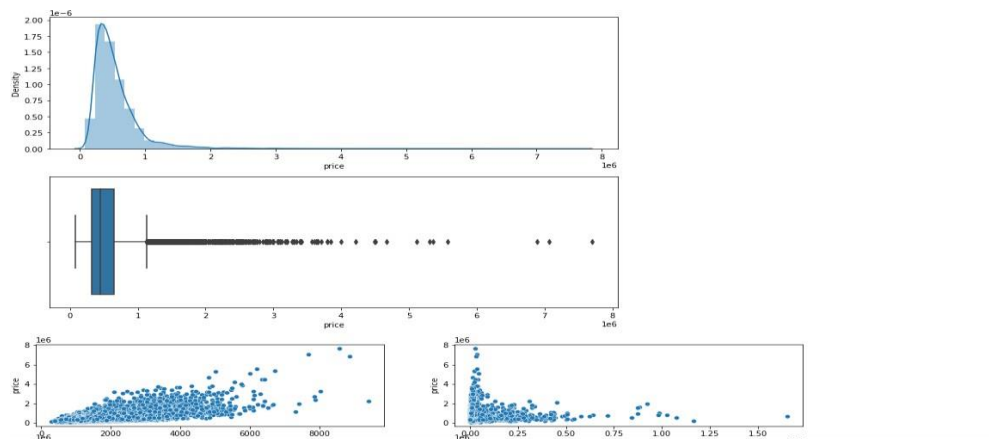


Figure 4.2 Visualizing house price

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.

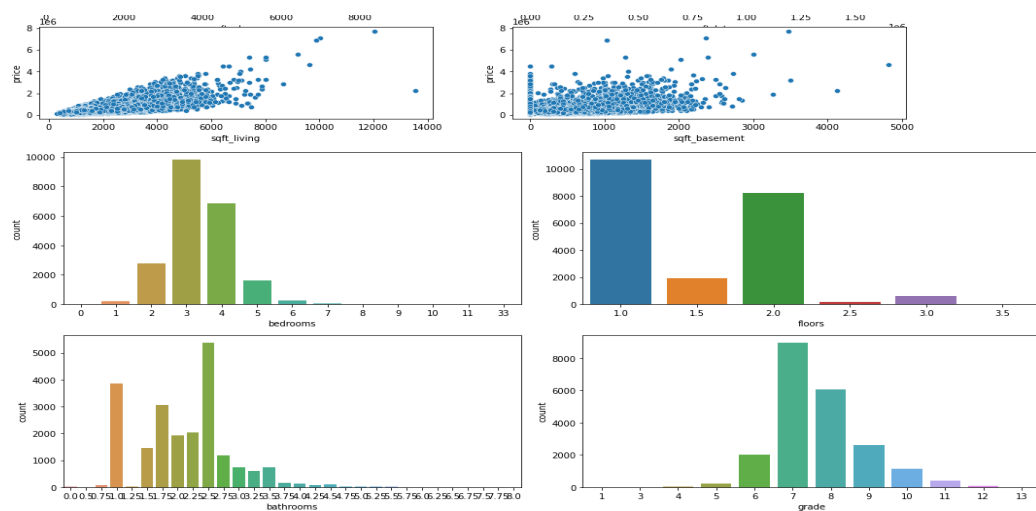


Figure 4.3 Count of bathroom, floor and bedrooms

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.

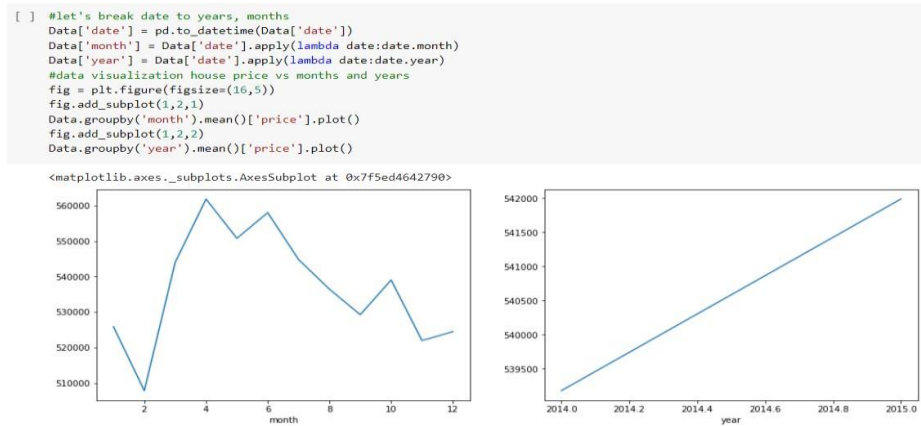


Figure 4.4 Price range month vs year

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 comprehensible 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] A. Ebekozen, A. R. Abdul-Aziz, and M. Jaafar, “Housing finance inaccessibility for low-income earners in Malaysia: Factors and solutions,” *Habitat Int.*, vol. 87, no. April, pp. 27–35, 2019, doi: 10.1016/j.habitatint.2019.03.009.
- [3] A. Jafari and R. Akhavian, “Driving forces for the US residential housing price: a predictive analysis,” *Built Environ. Proj. Asset Manag.*, vol. 9, no. 4, pp. 515–529, 2019, doi: 10.1108/BEPAM-07-2018-0100.
- [4] 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.
- [5] Phan TD. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering (ICMLDE) 2018*. doi:10.1109/icmlde.2018.00017.