A Minor Project Report

on

Hyderabad House price Forecast

submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

by

M. Manoj Kumar (20211A05G4)

M. Sravani (20211A05G2)

M. Koushik (20211A05F3)

Under the guidance of

Mrs. G. Vindhya M. Tech,
Assistant Professor.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

B.V.RAJU INSTITUTE OF TECHNOLOGY

(UGC Autonomous, Accredited by NBA & NAAC)

Vishnupur, Narspur, Medak(Dist.), Telangana State, India - 502313

2021 - 2022

B. V. Raju Institute of Technology
(UGC Autonomous, Accredited By NBA & NAAC)
Vishnupur, Narspur, Medak (Dist.),
Telangana State, India – 502313

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Minor Project entitled **"Hyderabad House price Forecast"**, being submitted by

M. Manoj Kumar (20211A05G4)

M. Sravani (20211A05G2)

M. Koushik (20211A05F3)

In partial fulfillment of the requirements for the award of degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING to B.V.RAJU INSTITUTE OF TECHNOLOGY is a record of bonafide work carried out during a period from May 2020 to July 2020 by them under the guidance of **Mrs. G. Vindhya**, Assistant Professor, CSE Department.

This is to certify that the above statement made by the students is/are correct to the best of my knowledge.

Mrs. G. Vindhya
Assistant Professor

The Project Viva-Voce Examination of this team has been held on _____.

Dr. Ch. Madhu Babu
Professor & HoD-CSE

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project entitled **"Hyderabad House price Forecast"** in partial fulfillment of the requirements for the award of Degree of Bachelor of Technology and submitted in the Department of Computer Science and Engineering, B. V. Raju Institute of Technology, Narsapur is an authentic record of my own work carried out during a period from May 2020 to July 2020 under the guidance of **Mrs. G. Vindhya**, Assistant Professor. The work presented in this project report has not been submitted by us for the award of any other degree of this or any other Institute/University.

M. Manoj Kumar (20211A05G4)
M. Sravani (20211A05G2)
M. Koushik (20211A05F3)

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely fortunate to have got this all along the completion. Whatever we have done is due to such guidance and assistance. We would not forget to thank them.

We thank **Mrs. G. Vindhya** for guiding us and providing all the support in completing this project. We are thankful to **Mrs. L. Pallavi**, our section project coordinator for supporting us in doing this project. I We are thankful to Mr. V. Pradeep Kumar, project coordinator for helping us in completing the project in time. We thank the person who has our utmost gratitude is **Dr. Ch. Madhu Babu**, Head of CSE Department.

We are thankful to and fortunate enough to get constant encouragement, support and guidance from all the staff members of CSE Department.

M. Manoj Kumar (20211A05G4)

M. Sravani (20211A05G2)

M. Koushik (20211A05F3)

HYDERABAD HOUSE PRICE PREDICTION

ABSTRACT

People are careful when they are trying to buy a new house with their budgets and market strategies. The objective of the model is to forecast the coherent house prices for non-house holders based on their financial provisions and their aspirations. By analysing the foregoing merchandise, fare ranges and also developments, speculated prices will be estimated. The model involves predictions using different Regression techniques like Multiple linear, Decision tree, random forest and rounding of using RSME technique. House price prediction on a data set has been done by using above mentioned techniques to find out the best among them. This model will also comprehensively validate multiple techniques in implementation regression and provide an optimistic result for housing price prediction.

The motive of this model 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 area size in sqft, number of rooms and location etc.

Key Words: House price prediction, Machine Learning, Linear regression, Decision tree, Random Forest.

TEAM MEMBERS:

1. Mamilla Manoj Kumar – 20211A5G4
2. Mamidgi Sravani – 20211A05G2
3. Madipadge Koushik – 20211A05F3

GUIDE:

Mrs. G. Vindhya – Assistant Professor

CONTENTS

List of Figures

List of Tables

List of Screens

List of Abbreviations

Fig – Figure

Chapter 1
INTRODUCTION

1.1 Motivation

This study aims to analyse the accuracy of predicting house prices when using linear, Lasso, Ridge regression algorithms. Thus, the purpose of this study is to deepen the knowledge in regression methods in machine learning. In addition, the given datasets should be processed to enhance performance, which is accomplished by identifying the necessary features by applying one of the selection methods

The main motive of this model are as follows:

• To apply data preprocessing and preparation techniques in order to obtain clean data

• To build machine learning models able to predict house price based on house features

• To analyze and compare models' performance in order to choose the best model

1.2 Problem Definition

People are careful when they are trying to buy a new house with their budgets and market strategies. The objective 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 developments, speculated prices will be estimated.

1.3 Objective of Project

The objective of this model 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 area size in sqft, number of rooms and location etc.

1.4 Limitations of Project

The local data will be requested from the Hyderabad. The request contains a list of features, that matches the public dataset's features, that is desired to be available when the data is sent. There is no guarantee that the data will be available in time nor contains the exact requested list of features. Thus, there might be a risk that the access will be denied or delayed. If so, the model will be accomplished based only on the public dataset. Moreover, this model will not cover all regression algorithms; instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced ones. Likewise, the artificial neural network that has many techniques and a wide area and several training methods that do not fit in this model.

1.5 Organization of Documentation

The organization of Documentation is as follows: Section 1 introduces the area of study. Section 2 gives an overview of the Literature. Section 3 gives an overview of the algorithms. shows the followed methods in this study, Section 4 presents the methods that are being used in the experiment in addition to the theoretical findings in addition to, the design of the experiment. Section 5 shows the experimental implementation process and the experiment results followed by a discussion in section 6. Finally, Section 7 concludes with remarks and hints about future work.

Chapter 2

LITERATURE SURVEY

2.1 Introduction

The performance will be measured upon predicting house prices since the prediction in many regression algorithms relies not only on a specific feature but on an unknown number of attributes that result in the value to be predicted. House prices depend on an individual house specification. Houses have a variant number of features that may not have the same cost due to its location. For instance, a big house may have a higher price if it is located in desirable rich area than being placed in a poor neighborhood. The data used in the experiment will be handled by using a combination of pre-processing methods to improve the prediction accuracy. In addition, some factors will be added to the local dataset in order to study the relationship between these factors and the sale price in Hyderabad

House Price prediction, is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improve Real Estate efficiency.

2.2 Existing System

The existing model for house price prediction uses Random Forest regression and takes only 6 parameters as data input for prediction of the house price which there by proves to less effective in predicting the close value.

2.3 Disadvantages of Existing system

- Low Accuracy as the system uses the outdated regression models

- Low features considered only few parameters are considered for predicting the value.

- Artificial data set the data set considered is the artificial data set which was made according to the mode attributes.

2.4 Proposed System

The present model enhances the previous version by adding the new regression technique LASSO and by inputting the more features effecting the price like car parking availability etc.

And overcoming the existing systems drawbacks by considering

Real time survey data, Maximum locations of Hyderabad, Adding XGboost for regression.

Chapter 3

ANALYSIS

3.1 Introduction:

Analysis is the process of dissecting something into its component elements in order to understand what each one does and how they interact with one another.

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriate in big data since attempting to manually process vast volumes of data would be impossible without the support of machines. Machine learning in computer science attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creating algorithms that permit the machine to learn. However, there are two general groups in machine learning which are supervised and unsupervised. Supervised is where the program gets trained on pre-determined set to be able to predict when a new data is given. Unsupervised is where the program tries to find the relationship and the hidden pattern between the data.

3.2 Software Requirement Specification

3.2.1 User requirements

Operating System: Windows 8/8.1/10, Linux and Mac

3.2.2 Software requirements

- Python 3.2 and above
- Pandas, Numpy
- Matplotlib

- Streamlit

3.2.3 Hardware requirments

• Processor: 2 gigahertz (GHz) or faster processor.
• RAM: 4 gigabytes (GB) for 32-bit or 8 GB for 64-bit.
• Hard disk space: =16GB.

RAM: 4 GB. If your PC has less than 4GB of memory, there are sometimes options for upgrading to get additional RAM. As the dataset contains with huge prices it requires 4GB and above.

3.3 Content Diagrams of Project

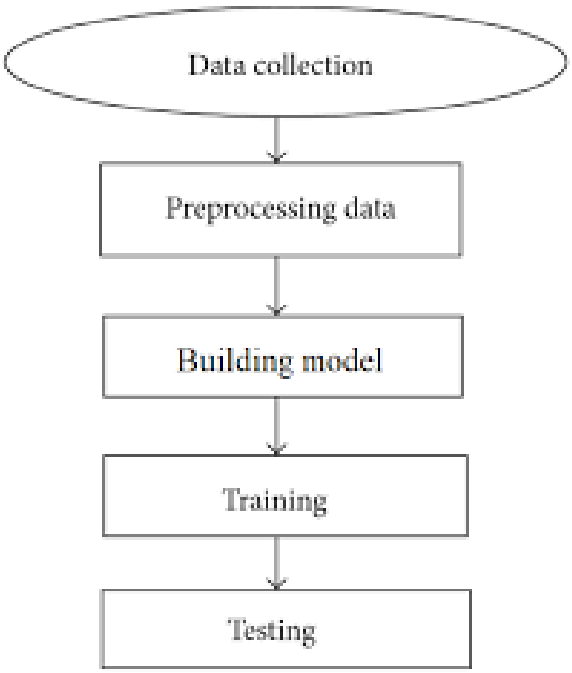


Fig. 3.3.1: Flowchart of prediction model

3.4 Algorithms and Flowcharts

3.4.1 Linear Regression:

Linear Regression is a supervised technique used to estimate the relationship between one dependent variable and more than one independent variables. Identifying the correlation and its cause-effect helps to make predictions by using these relations. To estimate these relationships, the prediction accuracy of the model is essential; the complexity of the model is of more interest. However, Linear Regression is prone to many problems such as multicollinearity, noises, and overfitting, which effect on the prediction accuracy. Regularised regression plays a significant part in Linear Regression because it helps to reduce variance at the cost of introducing some bias, avoid the overfitting problem and solve ordinary least squares (OLS) problems. There are two types of regularisation techniques L1 norm (least absolute deviations) and L2 norm (least squares). L1 and L2 have different cost functions regarding model complexity

3.4.2 Lasso Regression:

Least Absolute Shrinkage and Selection Operator (Lasso) is an L1-norm regularised regression technique that was formulated by Robert Tibshirani in 1996. Lasso is a powerful technique that performs regularisation and feature selection. Lasso introduces a bias term, but instead of squaring the slope like Ridge regression, the absolute value of the slope is added as a penalty term.

Lasso is defined as: $L = Min(sum\ of\ squared\ residuals + \alpha * |slope|)$ (1) Where $Min(sum\ of\ squared\ residuals)$ is the Least Squared Error, and $\alpha * |slope|$ is the penalty term. However, alpha α is the tuning parameter which controls the strength of the penalty term. In other words, the tuning parameter is the value of shrinkage. $|slope|$ is the sum of the absolute value of the coefficients

Cross-validation is a technique that is used to compare different machine learning algorithms in order to observe how these methods will perform in practice. Cross-validation method divides the data into blocks. Each block at a time will be used for testing by the algorithm, and the other blocks will be used for training the model. In the end,

the results will be summarised, and the block that performs best will be chosen as a testing block. However, α is determined 4 by using cross-validation. When α = 0, Lasso becomes Least Squared Error, and when $\alpha \neq 0$, the magnitudes are considered, and that leads to zero coefficients. However, there is a reverse relationship between alpha a and the upper bound of the sum of the coefficients t. When $t \to \infty$, the tuning parameter $a = 0$. Vice versa when $t = 0$ the coefficients shrink to zero and $a \to \infty$. Therefore, Lasso helps to assign zero weights to most redundant or irrelevant features in order to enhance the prediction accuracy and interpretability of the regression model. Throughout the process of features selection, the variables that still have non-zero coefficients after the shrinking process are selected to be part of the regression model. Therefore, Lasso is powerful when it comes to feature selection and reducing the overfitting.

3.4.3 Ridge Regression:

The Ridge Regression is an L2-norm regularised regression technique that was introduced by Hoerl in 1962. It is an estimation procedure to manage collinearity without removing variables from the regression model. In multiple linear regression, the multicollinearity is a common problem that leads least square estimation to be unbiased, and its variances are far from the correct value. Therefore, by adding a degree of bias to the regression model, Ridge Regression reduces the standard errors, and it shrinks the least square coefficients towards the origin of the parameter space. Ridge formula is: $R = Min(sum\ of\ squared\ residuals + \alpha * slope\ 2\ )$.Where $Min(sum\ of\ squared\ residuals)$ is the Least Squared Error, and $\alpha * slope$is the penalty term that Ridge adds to the Least Squared Error. When Least Squared Error determines the values of parameters, it minimises the sum of squared residuals. However, when Ridge determines the values of parameters, it reduces the sum of squared residuals. It adds a penalty term, where α determines the severity of the penalty and the length of the slope. In addition, increasing the α makes the slope asymptotically close to zero. Like Lasso, α is determined by applying the Cross-validation method. Therefore, Ridge helps to reduce variance by shrinking parameters and make the prediction less sensitive.

Chapter 4

DESIGN

4.1 INTRODUCTION

This section describes the working of the various modules in the model and the way in which they are communicating data with each other. This model is divided into two parts:

User Interface

- This is the page which displays all the data requirements and results of the data.

Back End

- In this all the analysis is done based on data set provided.

4.2 UML Diagrams

Sequence Diagram

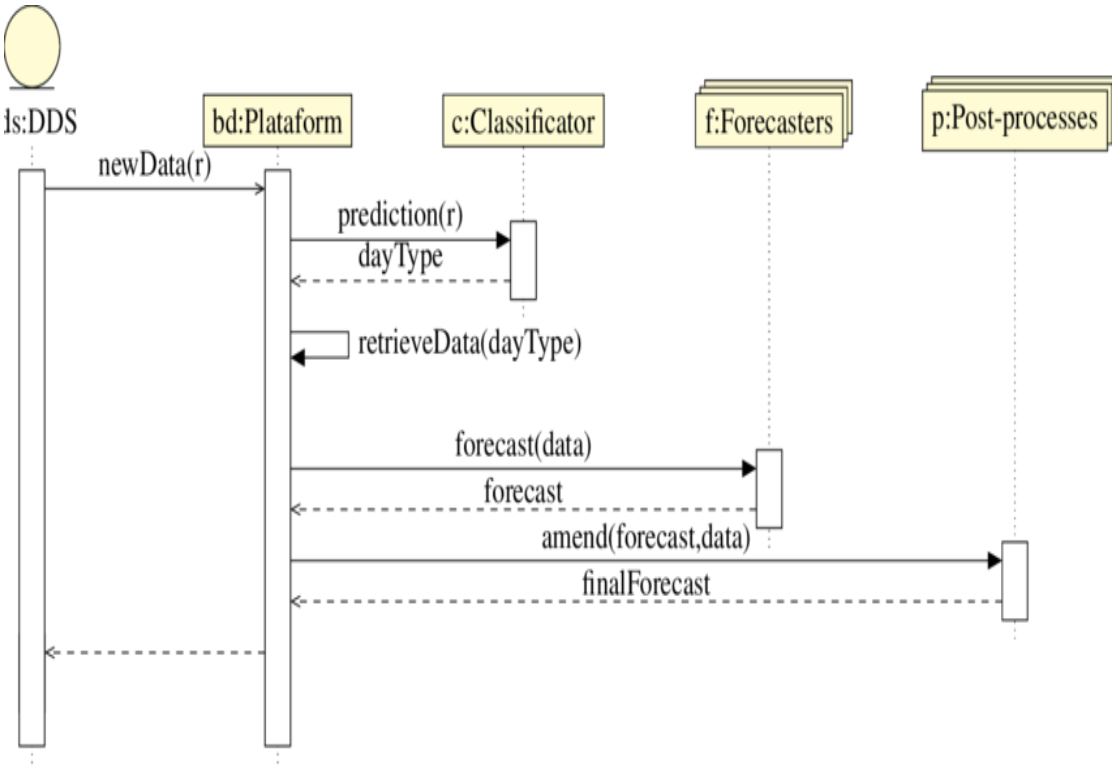


Fig. 4.2.1: Sequence diagram for house price prediction

DFD Diagrams

Level 0

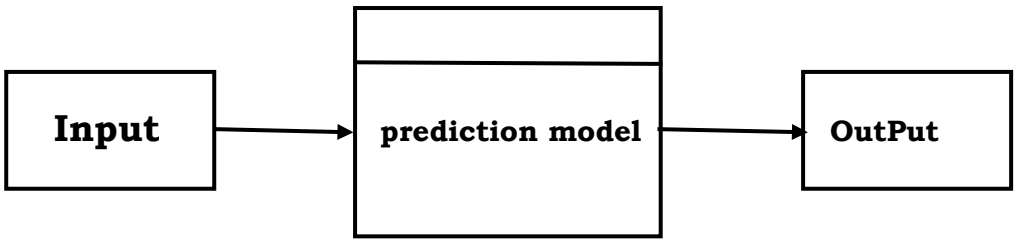


Fig. 4.2.2 DFD leve0 for house price prediction

Level 1

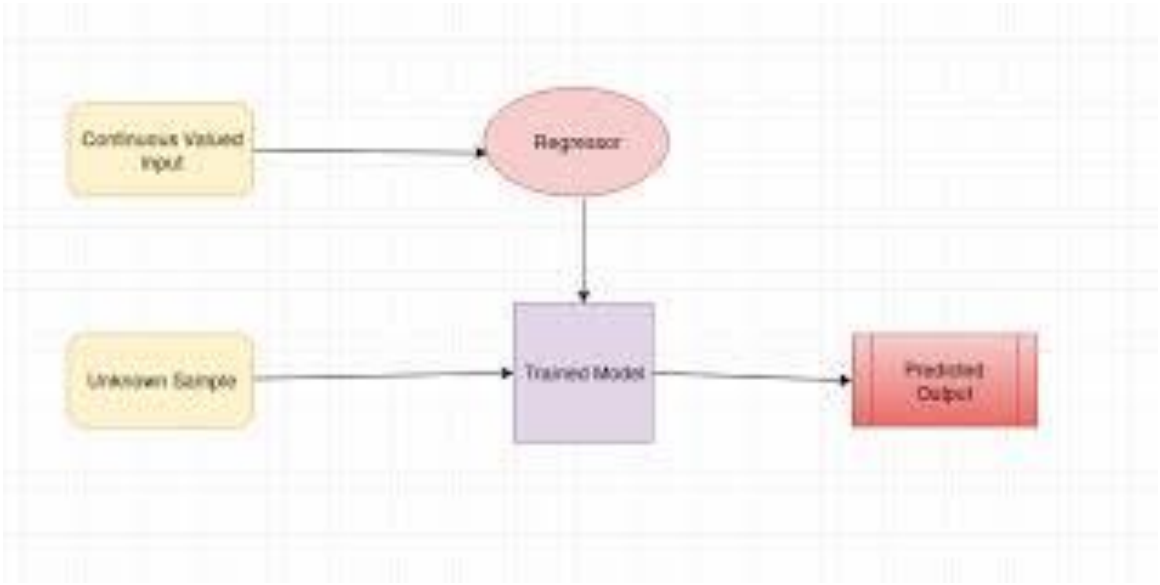


Fig. 4.2.3 DFD leve0 for house price prediction

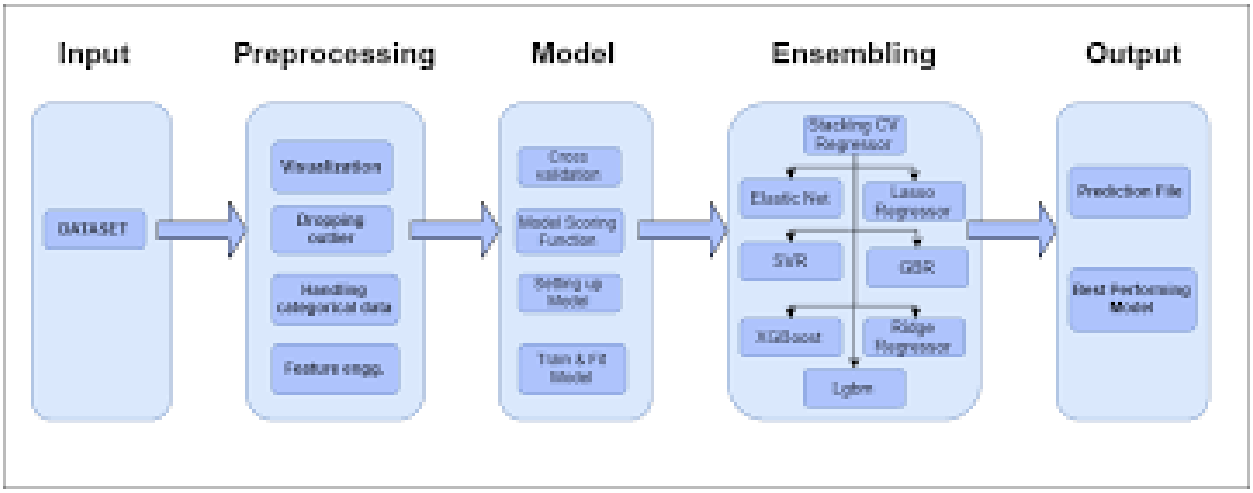


Fig. 4.2.4 DFD for house price prediction

Abstract view of model

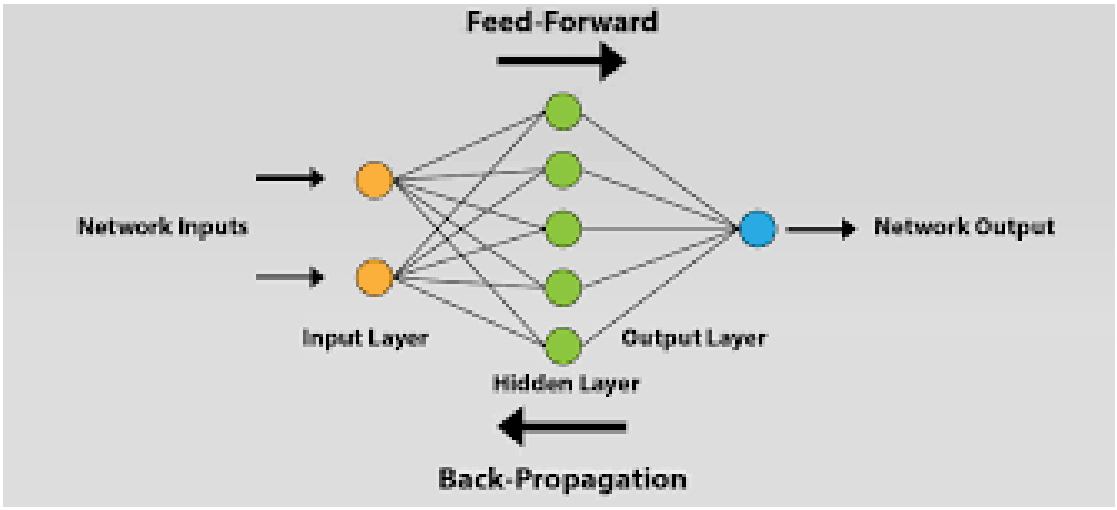


Fig. 4.2.5 abstract view of house price prediction model

4.3 Module design and organization

Data Collection

In this module, we collect different datasets using predefined libraries in Python.

Data Refining

The data may have some null values and words in which machine learning won't accept the characters. It only accepts numeric data.

Training and Testing

We split our data into training and testing data using a test split from the sci-kit learn library.

Finding accuracy

We find the accuracy score and correlation matrix for the test target variable to the predicted values. We have used sklearn to import the accuracy score.

Saving model

We have used the best model from the all the tested models and used the best one with highest accuracy for predicting the values.

Chapter 5

IMPLEMENTATION & RESULTS

5.1 INTRODUCTION

Implementation is one of the phases of the Software Development life Cycle. The implementation phase of software development involves the translation of design specifications source code and debugging, documentation and unit testing of the source code. In this project, we are going to develop a website which takes the input from users as features of the house. Based on the inputs we will predict the suitable crop to the farmer.

5.2 EXPLANATION OF KEY FUNCTIONS

- In this project, the streamlit module is used to create a webpage for the machine learning model.
- Using python, the main code of the model to predict crops is executed.
- The Numpy package is used to take the inputs in multi-dimensional array form.
- Pickle package is used to store the files in bytes form.

5.3 METHOD OF IMPLEMENTATION

5.3.1 Approach

In this project, the total code is divided into 3 parts.

Part 1 - trained_model

It contains the main code of the model about how to pre-process the dataset. Dataset is divided into a training dataset and testing dataset and algorithm classifiers are used to get the results.

	location	price	total_sqft	No. c	Schc	24X7	CarP	Hosp	LiftA	price_per_sqft
0	Adibatla	2800000	1400	3	0	0	0	0	0	200,000,000.0000
1	Adibatla	2800000	1400	3	0	0	0	0	0	200,000,000.0000
2	Adibatla	2022000	1010	2	0	0	0	0	0	200,198,019.8020
3	Adibatla	2200000	1050	2	0	0	0	0	0	209,523,809.5238
4	Adibatla	2800000	1200	2	0	0	0	0	0	233,333,333.3333
5	Adibatla	2200000	1000	2	0	0	0	0	0	220,000,000.0000
6	Adibatla	2500000	1100	2	0	0	0	0	0	227,272,727.2727
7	Adibatla	2304000	1002	2	0	0	0	0	0	229,940,119.7605
8	Adibatla	2300000	1000	2	0	0	0	0	0	230,000,000.0000
9	Adibatla	3742000	1134	2	0	1	1	0	1	329,982,363.3157

(1766, 8) (1766,) (197, 8) (197,)

Fig. 5.3.1.1: Training data overview

Part 2- predictive_system.py

Using Stream we will create a user interface to print the result of the input. In this, we will dump the main code of the model using the pickle model.

```
In [42]: import xgboost as xgb
         xgb1 = xgb.XGBRegressor()

In [43]: xgb1.fit(x_train,y_train)
         xgb_pred = xgb1.predict(x_test)

In [44]: xgb1.save_model("lasso.json")

In [45]: import streamlit as  st

In [46]: st.header("House price prediction")

         2022-07-20 22:15:16.460
           Warning: to view this Streamlit app on a browser, run it with the following
           command:

             streamlit run C:\Users\Mamilla Manoj\AppData\Local\Programs\Python\Python310\lib\site-packages\ipykernel_launcher.py [ARGUM
         ENTS]

Out[46]: DeltaGenerator(_root_container=0, _provided_cursor=None, _parent=None, _block_type=None, _form_data=None)

In [47]: np.save('classes.npy',le.classes_)

In [48]: le.classes_ = np.load('classes.npy',allow_pickle = True)

In [49]: xgb_best = xgb.XGBRegressor()

In [50]: xgb_best.load_model("lasso.json")
```

Fig. 5.3.1.2: Stream lit code for frontend

Part 3- webapp.py

This module takes .py module which is used to run the webpage. Stream lit module is used to run the webpage.

```
In [51]: if st.checkbox('Show Training Dataframe'):
             Data
```

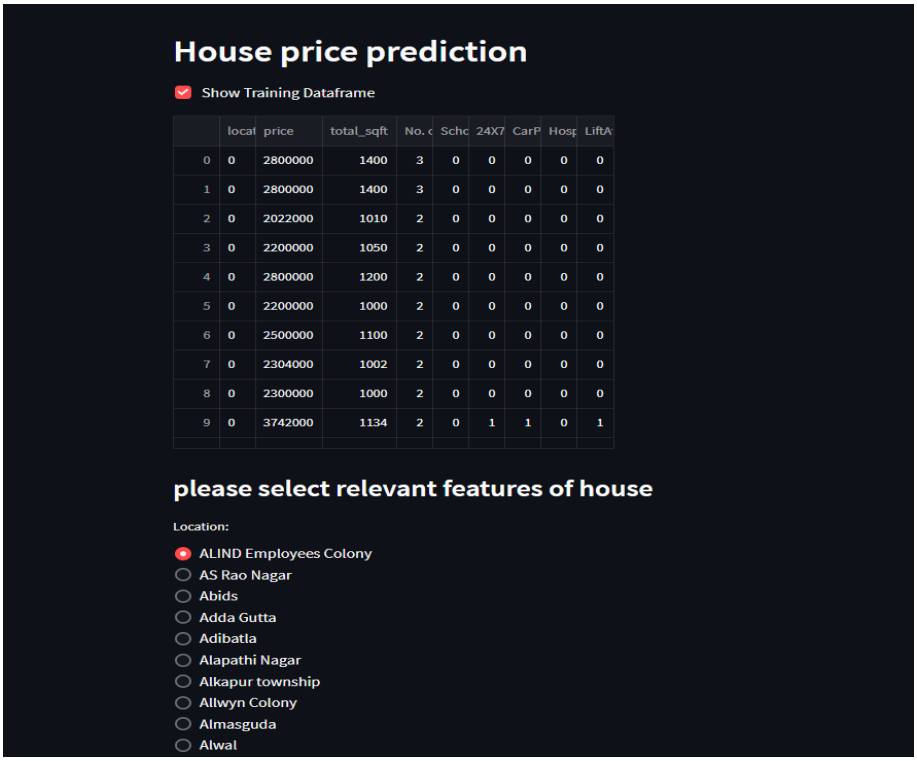
```
In [56]: st.subheader("please select relevant features of house")
         left_column, right_column = st.columns(2)
         with left_column:
             inp_species = st.radio('Location: ',np.unique(df['location']))
```

```
In [59]: input_sqft = st.slider('Total sqft(sqft)',0,max(Data["total_sqft"]))
         input_bed = st.slider('No.of bedrooms',0,max(Data["No. of Bedrooms"]))
         input_school = st.slider('school availabl',0,1)
         input_security = st.slider('24X7Security',0,1)
         input_car = st.slider('Car',0,1)
         input_hospital = st.slider('hospital',0,1)
         input_lift = st.slider('lift',0,1)
```

```
In [60]: if st.button('Make Prediction'):
             input_loc = le.transform(np.expand_dims(inp_species,-1))
             inputs = np.expand_dims(
                                 [int(input_species),input_sqft,input_bed,input_school,input_security,input_car,input_hospital,input_
             prediction = xbg_best.predict(inputs)
             print("final pred :",np.squeeze(prediction,-1))
             st.write(f"your house price: {np.squeeze(prediction, -1):.2f}g")
```

Fig 5.3.1.3: Deployment code for model

5.3.2 Output Screens



Screen 5.3.2.1: Web application screen

5.3.2 Result Analysis

Screen 5.3.2.1: Web application screen2

Chapter 6

TESTING AND VALIDATION

6.1 INTRODUCTION

Software testing serves as the final assessment of specification, design, and code generation and is a crucial component of software quality assurance. Testing is done to look for mistakes. Testing is the process of looking for any flaws or weaknesses in a piece of work. It offers a technique to test the efficiency of parts, subassemblies, and assemblies. It is the act of testing software to make sure it complies with user expectations and meets user needs without failing in an unacceptable way.

Software engineering testing methodologies are testing approaches, strategies, or techniques used to test a particular product to verify its usability. It ensures that the product performs as intended and has no negative impacts when utilised outside of its intended scope. Software testing strategies cover a wide range of testing techniques, including integration, security, and performance testing**.**

- Functional testing involves application testing in accordance with business requirements and a variety of test types intended to ensure that every component of the software acts exactly how the users would expect it to.
- Unit testing is a type of software testing that ensures that each individual piece of software is functioning flawlessly for the intended purpose at the code level.
- Integration Examination Each unit is rigorously tested before being merged with other units to produce modules or components that are intended to carry out particular functions.
- System testing is the black box testing method used to evaluate the integrated system as a whole and ensures it meets all specific requirements.

- Non-functional testing techniques use a variety of test kinds that are concentrated on a piece of software's operational features.
- The non-functional testing method known as performance testing is used to predict how an application will act in certain scenarios.
- Security evaluations Finding system flaws and security concerns is the aim of security testing.
- By evaluating a product with actual consumers, usability testing can determine how user-friendly a product is.

6.2 Design of test cases and scenarios

Using train_test_split method we have divide a single data set into training and testing data of desired lengths

Here we have taken 90 percent of data for training the model and 10 percent data for testing the model using test_size as 0.10 which indicates 10 percent of the data.

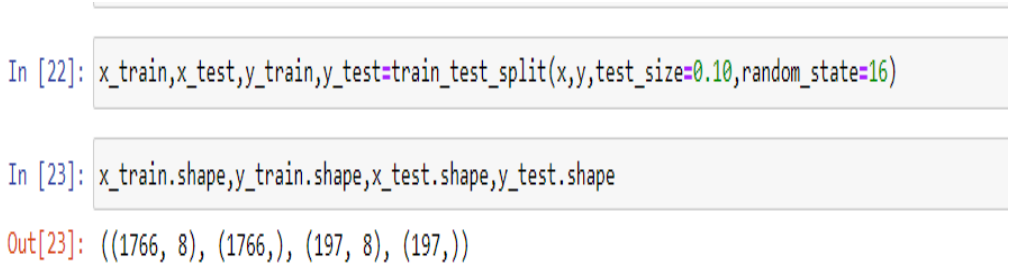


Fig 6.2.1: Code for splitting the training and testing data

6.3 Validation

Output:

Screen 6.3.1: Output displaying the predicted house price

When cross verified with the values in the available dataset, we came to know that the developed model predicted the outcome correctly.

Validation report

location	sqft	bedrooms	school	24x7 security	car	hospital	Expected Output	Result
Adibatla	1400	3	0	0	0	0	2800000	True
BHEL	1010	3	0	0	0	0	2022000	True
Alwal	1134	2	0	1	1	0	3724200	True
JUNTU	1350	3	0	1	0	1	5202000	False
BHEL	1100	2	2	0	0	0	2500000	True

Table 6.3.1: Expected output vs resultant output

Chapter 7

CONCLUSION AND FUTURE WORK

Conclusion:

- Machine learning serves to be one of the important components of the growing field of data science.
- By using statistical methods, algorithms are trained to make classifications predictions.
- Machine learning and data mining techniques are valuable in disease diagnosis. The ability to predict diabetes early serves a vital role in the patient's appropriate treatment procedure.
- We have subjected our dataset to various algorithms in this project and compared their respective accuracies.
- The study shows a comparison between the regression algorithms predicting house prices in Hyderabad. The results were promising for the public data due to it being rich with features and having strong correlation.
- Four machine learning algorithms were applied to the public Hyderabad house price dataset and trained and validated against a test dataset.
- Finally, we have realized that Lasso regression tends to have the maximum accuracy of 88 percent when compared with accuracies of linear regression (79.2%), ridge (75.3%).

Future Enhancement:

Future work on this model could be divided into two main areas to improve the result even further. Which can be done by:

- The used pre-processing methods do help in the prediction accuracy. However, experimenting with different combinations of pre-processing methods to achieve better prediction accuracy.
- Make use of the available features and if they could be combined as binning features has shown that the data got improved.
- The model will be deployed in the app so that could be available to every one to estimate the price of an house.

REFERENCES

Form Journal of Sweden University:

[1]. Ahmad Abdulal & Nawar aghi: “House Price Prediction”, Kristianstand university Sweden, pp. 3 – 28,, Issued on spring semester 2020 .

Form Book the Hundred-page Machine Learning Book:

[2]. Andriy Burkov: “The Hundred-page Machine Learning”, pp. 105, Quality South Asia Edition, 2006.

Form Research Papers:

[3]. I.J. Modern Education and Co mputer Science, 2020, 6, 46-54 Published Online December 2020 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijmecs.2020.06.04

Form Publication of IJRASET:

[4]. Authors: Mr. Piyush Chordia, Mr. Pratik Konde, Ms. Supriya Jadhav, Hrutik Pandhare, Prof. Shikha Pachouly DOI Link: https://doi.org/10.22214/ijraset.2022.40466

Git resource used:

https://github.com/Rohit7594/Banglore-House-Price-Predictor