**REAL ESTATE PRICE PREDICTION**

(Machine Learning Web App Made With Flask)

**PROJECT REPORT**

***by***

**EKAAGRA DUBEY**

****

# ABSTRACT

The below document presents the implementation of price prediction project

for the real estate markets and housing. Many algorithms like linear regression, XGboost, lasso regression, decision tree regressor, etc. are used here to effectively increase the accuracy percentage, various researchers have done this project

and implemented the algorithms like linear regression, Lasso regression,

Gradient Boost, decision tree which is considered as the best models in the price prediction.

These are considered as the base models and by the help of advanced data mining

tools algorithms like a gradient boosted trees and ensemble learning models are used and prediction accuracy is attained in a higher rate. Visualization techniques like count plot, distribution plot and box plots will help in pre-processing of the dataset. The results and evaluation of these models using the machine learning techniques like MSE, R^2 error analysis and visualizations like distribution plot and error difference plot help identify mistakes n fit of the model, which further help in giving a better fit and accuracy of the model

Keywords:,Multiple Regression,lasso regression,Gradient

boosted trees, Machine learning,Flask.

# 

# ACKNOWLEDGEMENT

I express my deep gratitude and regards to my instructors **Rav Ahuja**, **Alex Aklson**, **Romeo Kienzler**, **Polong Lin**, **Svetlana Levitan**, **Joseph Santarcangelo**, **Hima Vasudevan**, **Saeed Aghabozorgi** and to the rest **cousera team** & IBM educators for their valuable guidance ,monitoring ,encouragement and suggestions throughout my project work. I am very thankful they helped me learn skill like data science ,ML, Python & SQL, analyze & visualize data, build machine learning models.

I would also like to thank the platform **cognitiveclass.ai** powered by **IBM Developer Skills Network** to provide me with valuable and critical resources required in this project

At last I **Dr. AKHILESH DAS GUPTA ISTITUTE OF TECNOLOGY AND MANAGEMENT** to give me this opportunity for working on this project to enhance my skills.

# TABLE OF CONTENTS

**Certificate of Training ii**

**Abstract iii**

**Acknowledgement iv**

**Table of Content v**

**List of Figures vii**

**List of Tables viii**

**Chapter 1: Introduction 1**

* 1. Introduction to project 1

**Chapter 2: Related Work 2**

2.1 Introduction 2

2.2 Summary of literature review 2

**Chapter 3: Data Science 3-4**

3.1 Data science 3

3.2 Machine learning 3

3.3 Deep learning 3

3.4 Artificial intelligence 3

3.5 Relation 3-4

**Chapter 4: Methodology 5-10**

4.1 Description of dataset 5-6

4.2 Data cleaning & integration 6-8

4.2.1 Inconsistent column 6

4.2.2 Missing values 6

4.2.3 Outliers 7

4.2.4 Duplicate rows 7

4.2.5 Tidy data set 7

4.2.6 Converting data types 7

4.2.7 String manipulation 8

4.2.8 Data Concatenation 8

4.2.9 Data Visualization 8

4.3 Machine learning algorithms 8-10

4.3.1 Multiple regression 8-9

4.3.2 Cost function 9

4.3.3 Lasso regression 9

4.3.4 Gradient boosting algorithm 9-10

**Chapter 5: Implementation 11-20**

5.1 Cleaning dataset 11-17

5.2 Feature engineering 17-18

5.3 Feature normalization 19

5.4 Fitting model 19-20

5.4.1 Separating variables 19

5.4.2 Grid search 20

**Chapter 6: Results 21**

**Chapter 7 Model deployment 22-24**

7.1 Flask implementation 22-23

7.2 Web app interface 24

**Chapter 8: Conclusion 25**

8.2 Conclusion 25

8.3 Further Enhancements 25

**Reference 26**

**LIST** **OF** **FIGURES**

|  |  |
| --- | --- |
| Figure 3.1: | Relationship b/w DS,ML,AI &DL |
| Figure 4.1: | Project timeline |
| Figure 4.2: | Dataset description code |
| Figure 5.1: | NaN visualization |
| Figure 5.2: | Skewness description code |
| Figure 5.3: | Before removing outliers |
| Figure 5.4: | After removing outliers |
| Figure 5.5: | Visualization after removing outliers |
| Figure 5.6:  Figure 6.1  Figure 6.2 | Location count  Similar Price map  Predicted & Original price plot |
| Figure 7.1: | Flask implementation code |
| Figure 7.2: | Web app interface |

**CHAPTER 1: INTRODUCTION**

**1.1 Introduction :**

We need a proper prediction on the real estate and the houses in housing market we can see a mechanism that runs throughout the properties buying and selling buying a house will be a life time goal for most of the individual but There are lot of people making huge mistakes in India right now when buying the properties most of the people are buying properties unseen from the people they don’t know by seeing the advertisements and all over the grooves coming around the India one of the common mistakes is buying the properties that are too expensive but its not worth it. In the housing market .There was an economic collapse in the year 2007 and 2008 so there were several economic indicators that give the clue of impending disaster, this situation is currently happening and the economic indicators suggest that the housing prices are getting high people uses the real estate to known the current economic situations, the Indian government fails to produce the data about the house prices so it becomes difficult to buy the properties so the 87% of the people who needs to buy houses are using the Internets to search so there is evidence there is a correlation between housing sales and housing prices. In general, real estate may have the valuation of land may be obliged to furnish.

A quantitative measure of the problem was carried out by many different Players in the commercial centre, for example, land agents, Appraisers, assessors, mortarboard lenders, brokers, Developers, gurus Also reserve managers, lenders ,etc. Business worth will be evaluated through that requisition. From claiming valuation systems Also methods that reect those nature of property and the condition under which those provided for. The property might well on the way exchange in open market under many conditions and circumstances, people are the unaware amount the current situations and they start losing their money, the change in prices of properties would affect both the common people and the government, to avoid certain circumstances there is a need of price prediction.

Many methods have been used in the price prediction like a hedonic regression in this I am trying to predict the predict the real estate price for the future using the machine learning techniques with the help of the previous works. I have used the multiple regression and more algorithms with different tools to predict the house price So, it would be helpful for the people, so they will aware of both current and future situations, so it may avoid them in making mistakes. The remaining paper is organized as section 2 describes the previous works done by different researchers using different algorithms section 3 provides the methodology and the tools used and section 4 explains the way that the algorithms implemented, and comparisons and results are given in the last section.

# CHAPTER 2: RELATED WORK

**2.1 Introduction**

Before committing to the project several ground works should be done so there is a need for literature review I have analyzed many papers regarding the price prediction related the house markets and other different sectors. The papers I have taken will be in the different years up to the present year and I have used the recent and latest technologies, our main goal is to get more accuracy than the previous works the below passages will describe the past prediction works done by the various researchers and it will be helpful to implement the corresponding project.

## 2.2 Summary of Literature Review

For every project the literature review will give clear idea and it will serve as the base line

here most of the authors have concluded that artificial neural networks have the more

influence in predicting but in the real world the other algorithms should be also taken

into consideration. by conducting this study it helps to know about both the pros and

cons and it had helped me to successfully implement the project.

## 

# CHAPTER 3: DATA SCIENCE

# 

## 3.1 Data Science:

# Data science refers to the process of extraction of useful insights from data. This interdisciplinary approach merges various fields of computer science, scientific processes and methods, and statistics in order to extract data in automated ways.

# In order to mine big data, which is closely associated with the field, data science uses a diverse range of techniques, tools and algorithms gleaned from the fields. Data science training advances these techniques.

## 3.2 Machine Learning:

The field focuses on letting algorithms learn from the provided data, collect insights, and make predictions on unanalyzed data based on the gathered information. In general, machine learning is based on three key models of learning algorithms:

* supervised [machine learning algorithms](https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)
* unsupervised machine learning algorithms
* reinforcement machine learning algorithms.

In the first model, a dataset is present with inputs and known outputs. In the second one, the machine learns from a dataset that comes with input variables only. In reinforcement learning model, algorithms are used to select an action.

**3.3 deep learning**

Deep learning is an AI function that mimics the workings of the human brain in processing data for use in detecting objects, recognizing speech, translating languages, and making decisions. Deep learning AI is able to learn without human supervision, drawing from data that is both unstructured and unlabeled

.

## 3.4 Artificial Intelligence :

Though*it’s a broad term, at its core,****artificial intelligence (AI)****refers to the process of making machines enable to simulate the human brain function.*

In the modern technology landscape, artificial intelligence is divided into two key areas.

The first one is general AI, which is based on the concept that a system can handle tasks like speaking and translating, recognizing sounds and objects, performing business or social transactions etc. The other one is applied AI that refers to concepts like driverless cars.

## 3.5 Relation :

*The* interdisciplinary field of data science uses key skills of a wide range of fields including machine learning, statistics, visualization etc. It enables us to identify meaning and appropriate information from huge volumes of data to make informed decisions in technology, science, business etc.

For a simpler view on the relation between these technologies, ***artificial intelligence is applied based on machine learning***. And machine learning is a part of data science that draws features from algorithms and statistics to work on the data extracted from and produced by multiple resources. Thus, you can say data science merges together a bunch of algorithms obtained from

machine learning to develop a solution, and during the process, lots of ideas from traditional domain expertise, statistics and mathematics are borrowed.

# Significance of Data Science with A.I and M.L As stated in the above figure 2. The Data science field will make use of A.I algorithms and machine learning in order to make the effective and useful decisions. These decisions will be based on the user choices that how they need their data presentations like statistical, pictorial, textual and any other form. These representation of data is directly proportional with data processing by using Machine learning and A.I algorithms. These algorithms applied by using statistical, analytical and mathematical approaches.

Figure 3..: Relationship b/w DS,ML,AI &DL

# CHAPTER 4: METHODOLOGY

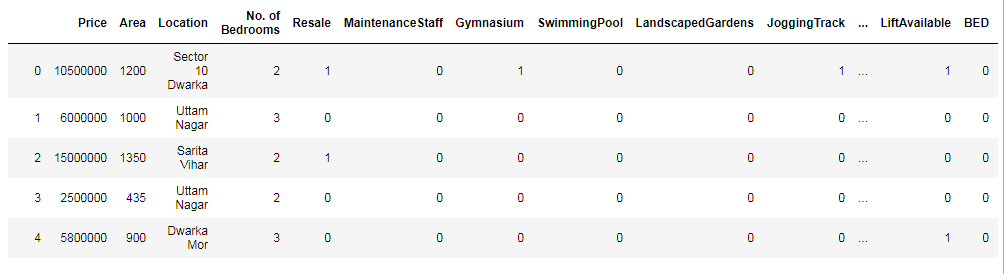
## 

Figure 4..: Project time line

## 

## 4.1 Description of Dataset:

The data used in this project is taken from kaggle. It contains about 4000 plus rows and multiple columns. The view and description of the data is given bellow



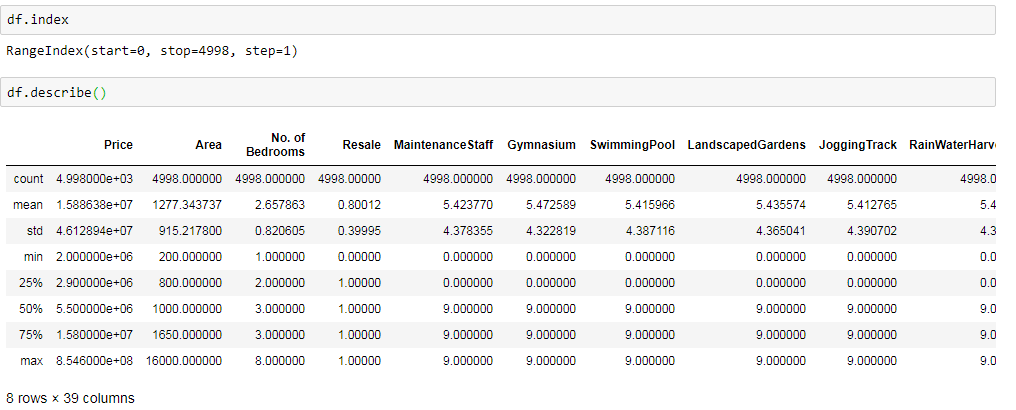


Figure 4.2: Dataset description code

**4.2 Data Cleaning and integration:**

Data Cleaning means the process of identifying the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying, replacing or deleting them according to the necessity. Data cleaning is considered a foundational element of the basic data science.

Data is the most valuable thing for Analytics and Machine learning. In computing or Business data is needed everywhere. When it comes to the real world data, it is not improbable that data may contain incomplete, inconsistent or missing values. If the data is corrupted then it may hinder the process or provide inaccurate results.

**Different ways of cleaning data:**

**4.2.1 Inconsistent column :**

If your DataFrame (A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns) contains columns that are irrelevant or you are never going to use them then you can drop them to give more focus on the columns you will work on.

**4.2.2 Missing data:**

It is rare to have a real world dataset without having any missing values. When you start to work with real world data, you will find that most of the dataset contains missing values. Handling missing values is very important because if you leave the missing values as it is, it may affect your analysis and machine learning models. So, you need to be sure that whether your dataset contains missing values or not. If you find missing values in your dataset you must handle it. If you find any missing values in the dataset you can perform any of these three task on it:

* drop corresponding row or column
* Replace it
* Leave as it is

**4.2.3 Outliers:**

If you are new data Science then the first question that will arise in your head is “what does these outliers mean” ? Let’s talk about the outliers first and then we will talk about the detection of these outliers in the dataset and what will we do after detecting the outliers.  
According to wikipedia,  
“*In statistics, an****outlier****is a data point that differs significantly from other observations.*”  
That means an outlier indicates a data point that is significantly different from the other data points in the data set. Outliers can be created due to the errors in the experiments or the variability in the measurements.

An outlier is an extremely high or extremely low-value value in the data it can be identi\_ed

if whether the value is greater than interquartile range Q3 + 1.5 or Q1 - 1.5 detecting the

interquartile range is arrange the data in an order from the lower value to the higher value,

now the mean is taken for the \_rst set of values and second set values now by subtracting

both mean we can get the interquartile range the formula is Q3 + (1.5)(quartile range)

and for Q1-(1.5)(quartile range)

For detecting the outliers we can use :  
1. Box Plot  
2. Scatter plot  
3. Z-score etc**.**

**4.2.4 Duplicate rows:**

Datasets may contain duplicate entries. It is one of the most easiest task to delete duplicate rows. To delete the duplicate rows you can use —  
dataset\_name.drop\_duplicates().

**4.2.5 Tidy data set:**

Tidy dataset means each columns represent separate variables and each rows represent individual observations. But in untidy data each columns represent values but not the variables. Tidy data is useful to fix common data problem.You can turn the untidy data to tidy data by using [pandas.melt](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.melt.html).

**4.2.6 Converting data types:**

In DataFrame data can be of many types. As example :  
1. Categorical data  
2. Object data  
3. Numeric data  
4. Boolean data

Some columns data type can be changed due to some reason or have inconsistent data type. You can convert from one data type to another by using [pandas.DataFrame.astype](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.astype.html)

**4.2.7 String manipulation:**

One of the most important and interesting part of data cleaning is string manipulation. In the real world most of the data are unstructured data. String manipulation means the process of changing, parsing, matching or analyzing strings. For string manipulation, you should have some knowledge about regular expressions. Sometimes you need to extract some value from a large sentence. Here string manipulation gives us a strong benefit. Let say,  
“This umbrella costs $12 and he took this money from his mother.”  
If you want to exact the “$12” information from the sentence then you have to build a regular expression for matching that pattern.After that you can use the python libraries.There are many built in and external libraries in python for string manipulation.

**4.2.8 Data Concatenation:**

In this modern era of data science the volume of data is increasing day by day. Due to the large number of volume of data data may stored in separated files. If you work with multiple files then you can concatenate them for simplicity. You can use the following python library for concatenate***.***

**4.2.9 Data visualization**

Data 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..

**4.3 Machine Learning Algorithms**

A. Linear regression: Simple linear regression statistical method allows us to summarize and study the relationship between two continuous quantative variables. One variable, denoted x, is regarded as the• predictor, explanatory, or independent variable. The other variable, denoted y, is regarded as• the response, outcome, or dependent variable.

**4.3.1 Multiple Regression:**

Analysis Multiple regression analysis is used to check whether there is a statistically noteworthy association the middle of sets of variables. It’s used to discoverpatterns in the individuals sets of information. Numerousrelapse Investigation will bevery nearlythe same Likewise basic straight relapse. The main distinction the middle of straightforward straight relapse Also numerous relapse is in the number for predictors (“x” variables) utilized within those relapse. Straightforward relapse examination employments An absolute x variable to each subordinate “y” variable. Case in point: (x1, Y1). Numerous relapse utilization numerous “x” variables for every free variable: (x1)1, (x2)1,(x3)1, Y1). In one-variable straight regression, you might information particular case subordinate variable (i. E. “sales”) against a autonomous variable (i. E. “profit”). Anyhow you could make intrigued by how diverse sorts from claiming offers impact the relapse. You Might set your X1 as particular case kind from claiming sales, your X2 Similarly as in turn sort about deals etc.

**4.3.2 The cost Function:**

Thus let’s say, you expanded the size of a specific shop, the place you predicted that those deals might a chance to be higher. Be that in spite of expanding those size, those bargains in that shop didn't expand that a great deal. Thereabouts those expense connected Previously, expanding those span of the shop, provided for you negative outcomes. So, we necessity on minimize these cost. So we present an expense function, which is fundamentally used to characterize and measure those slip of the model.

**4.3.3 Lasso Regression:**

Lasso regression which may be a standout among those relapse models that would accessible will examine the information. Further, the regression model may be demonstrated for a sample and the formula is Additionally recorded to reference. LASSO stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a standout among the regularization routines that makes niggardly models in the vicinity for vast number for features, the place expansive implies whichever of the following two things:. • Vast enough to improve those inclination of the model on over-fit. Least ten variables can foundation over fitting. • Huge enough will cause computational tests. This circumstance could emerge in the event from claiming a large number or billions about Characteristics. Tether relapse performs L1 regularization that is it includes those punishment equal of the supreme esteem of the extent of the coefficients. Here the minimization goal will be Concerning illustration emulated. Minimization goal = LS Obj + λ (sum about outright esteem of coefficients). The place LS Obj remains for minimum squares objective which will be nothing yet the straight relapse target without regularization Furthermore λ may be those turning figure that controls the measure for regularization. The inclination will build with those expanding quality of λ and the difference will diminish Concerning illustration the measure for shrinkage (λ) increments. The lasso regression estimate is defined as Here the turning component λ controls those quality for penalty, that is. When λ = 0: we get same coefficients Similarly as basic straight relapse. At λ = ∞: constantly on coefficients are zero. The point when 0 < λ < ∞: we get coefficients between 0 What's more that for basic straight relapse thus At λ is amidst the two extremes, we would adjusting those underneath two plans. Fitting An straight model for y once X.• Contracting those coefficients.•

**4.3.4 Gradient Boosting algorithm.:**

Gradient boosting is a machine Taking in strategy to relapse Also arrangement problems, that produces a prediction model in the structure of an group from claiming powerless prediction models. The exactness of a predictive model might be helped to two ways:. Possibly by graspingcharacteristic building alternately. Toward applyingboosting calculations straight far. There are a significantnumber boosting calculationsin.

* Gradient Boosting
* XGBoost
* AdaBoost
* Gentle Boost

Each boosting algorithm need its own underlying math. Also, a slight variety may be watched same time applying them. Boosting calculation will be a standout among those The greater part capable Taking in thoughts acquainted in the final one twenty A long time. It might have been intended to order problems, yet all the it can be developed should relapse too. The inspiration to gradient boosting might have been An technique. That combines those outputs about large portions “weak” classifiers to process An capable “committee. ” a powerless classifier (e. G. Choice tree) will be person whose slip rate is main superior to irregular guessing.

# CHAPTER 5 : IMPLIMENTATION

**5.1 Cleaning Dataset:**

Removing NaN values:

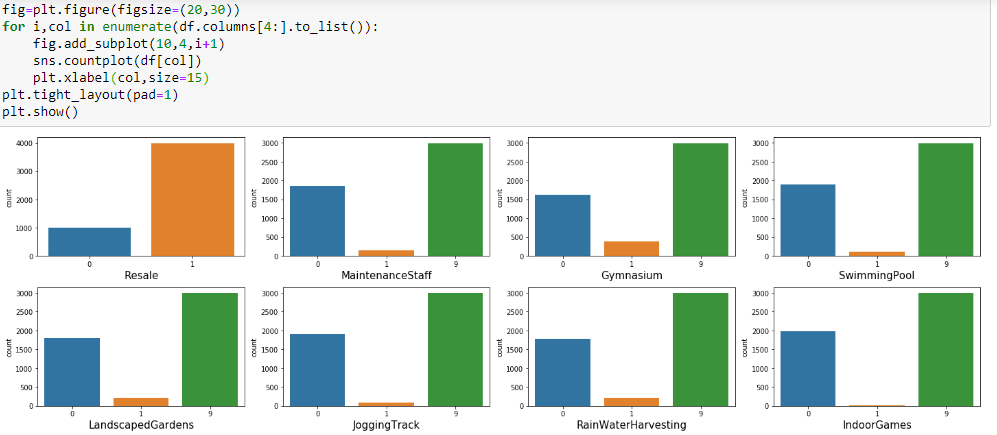
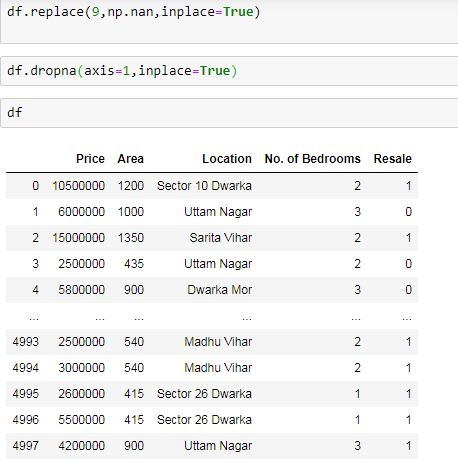
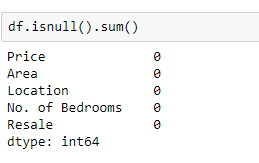
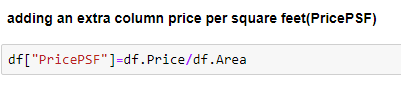


Figure 5..: NaN visualization

Since there are more 9(NaN) values than 0,1 we will be removing all the corresponding columns







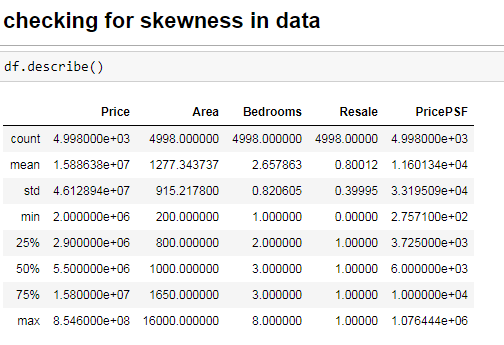
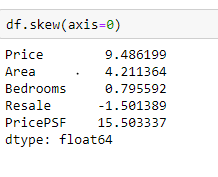
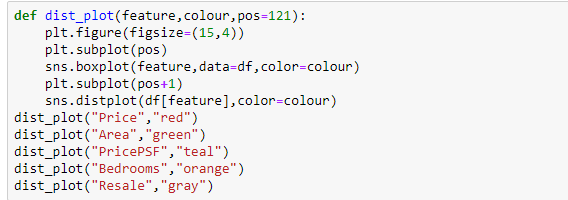
****

Figure 5.2: Skewness description code

****

Plotting skewness:



Result of plot:

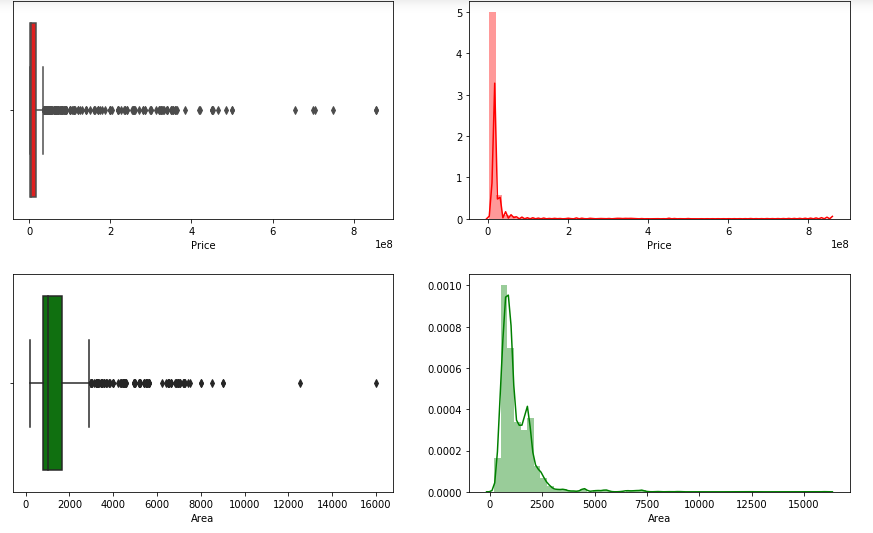
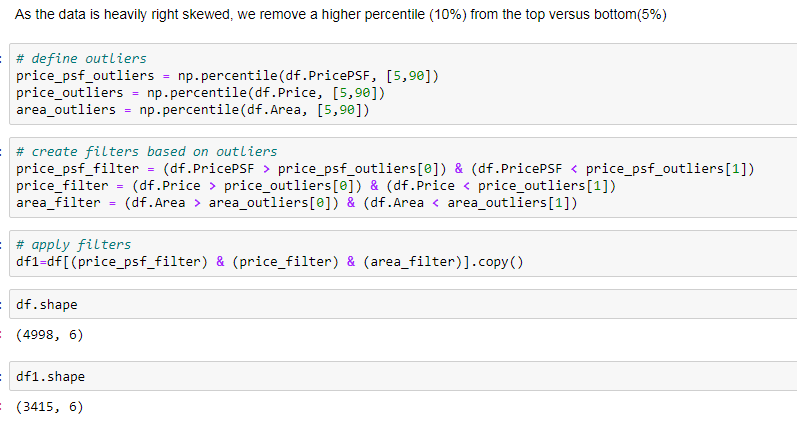


Figure 5.3: Before removing outliers

Code to remove outliers:



Plot after removing outliers:

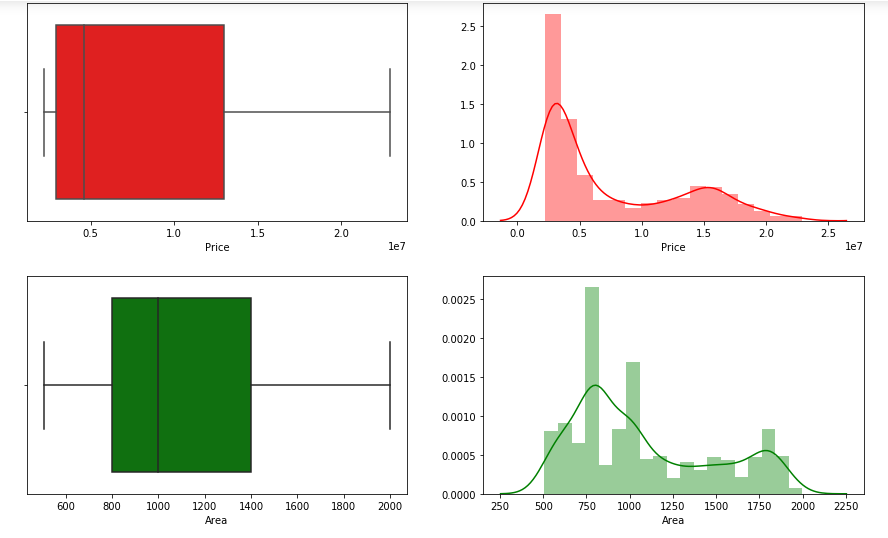


Figure 5.4: After removing outliers

Visualizations:

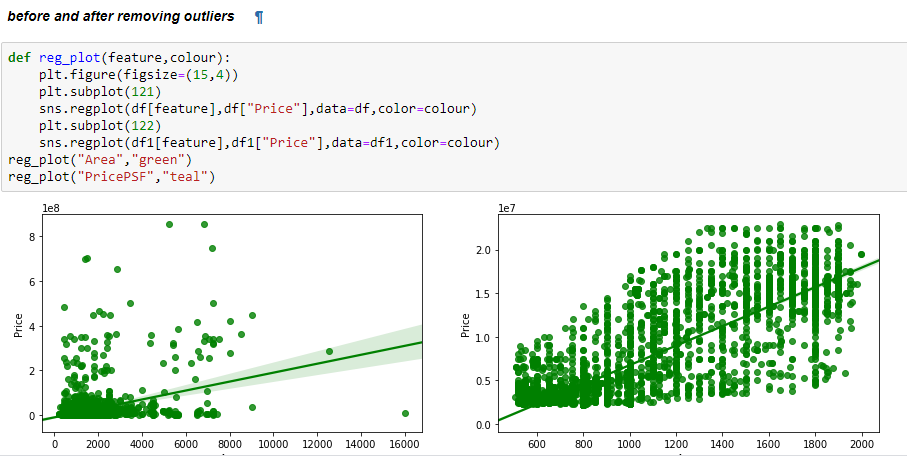
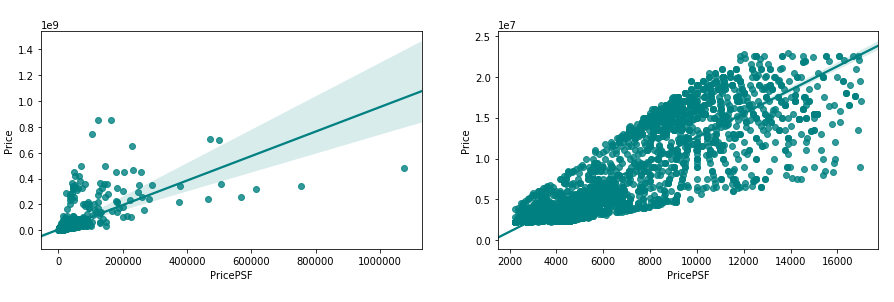
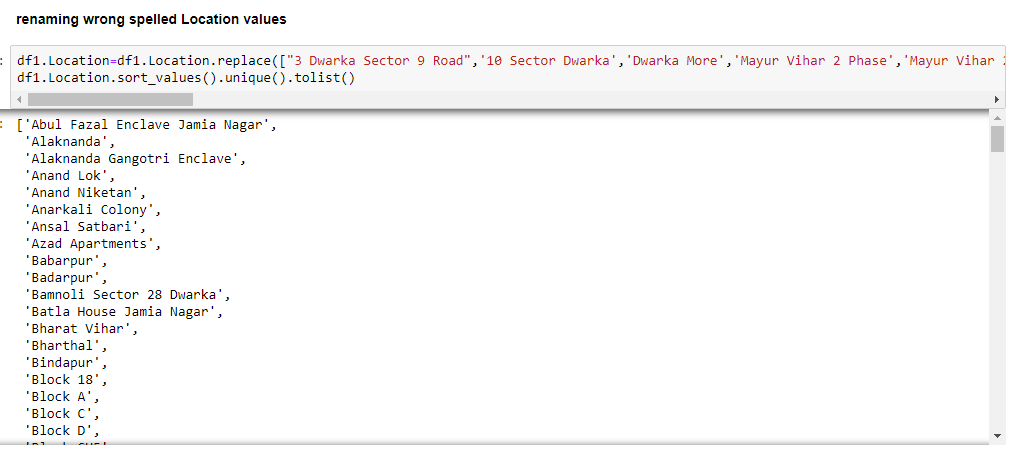
 

Figure 5.5 Visualizations after removing outliers



Plot to check count of each location:

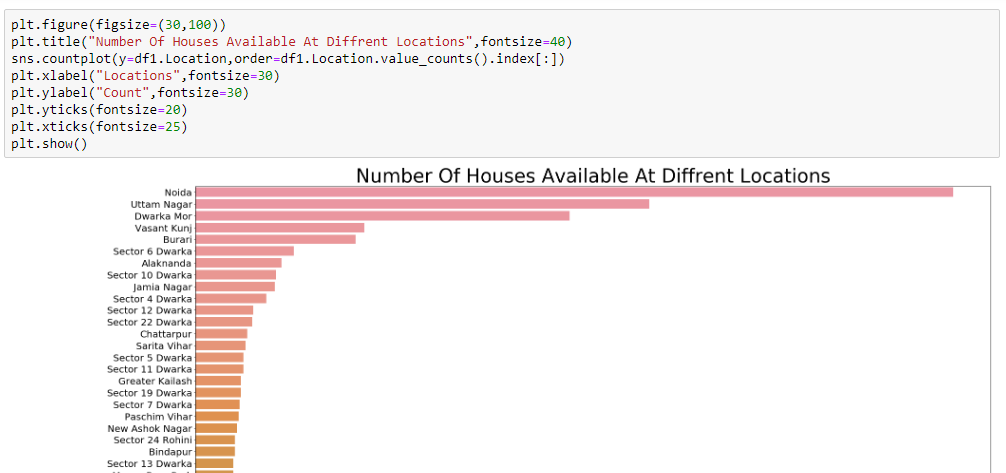
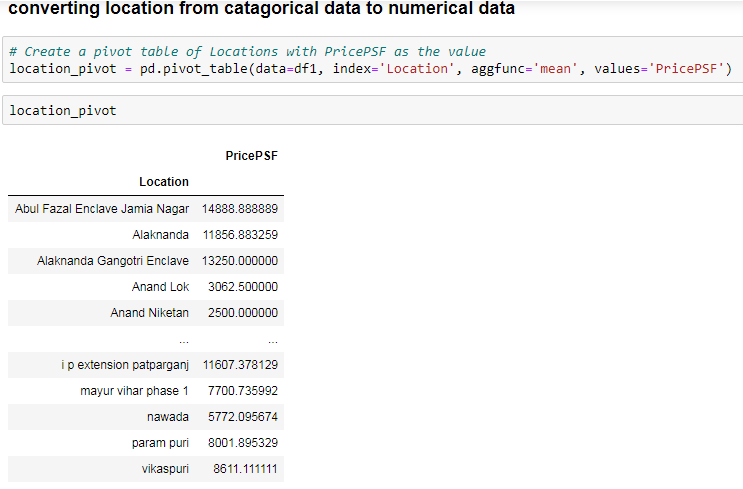
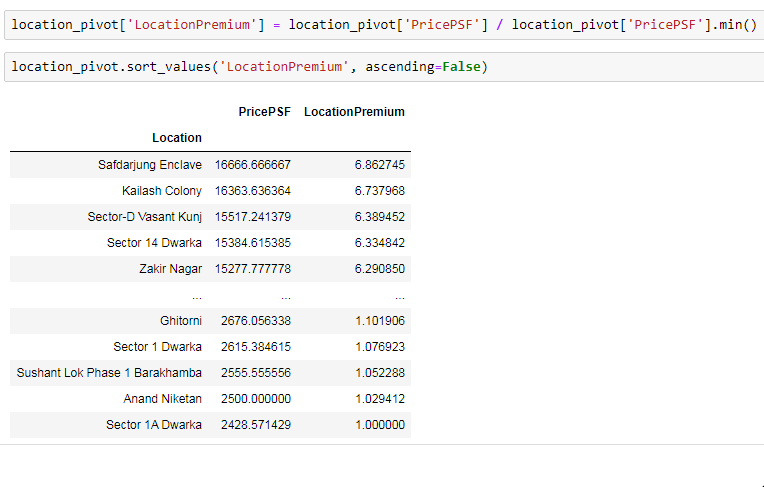
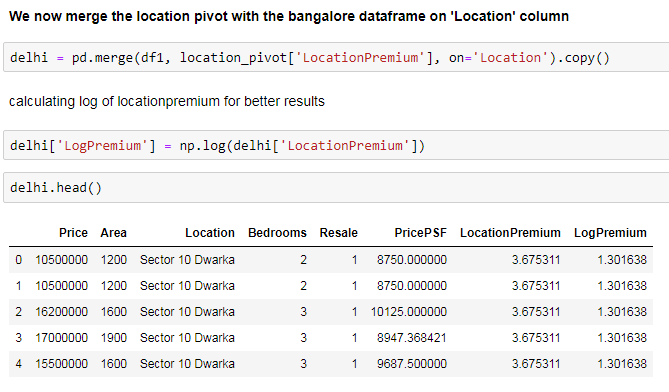


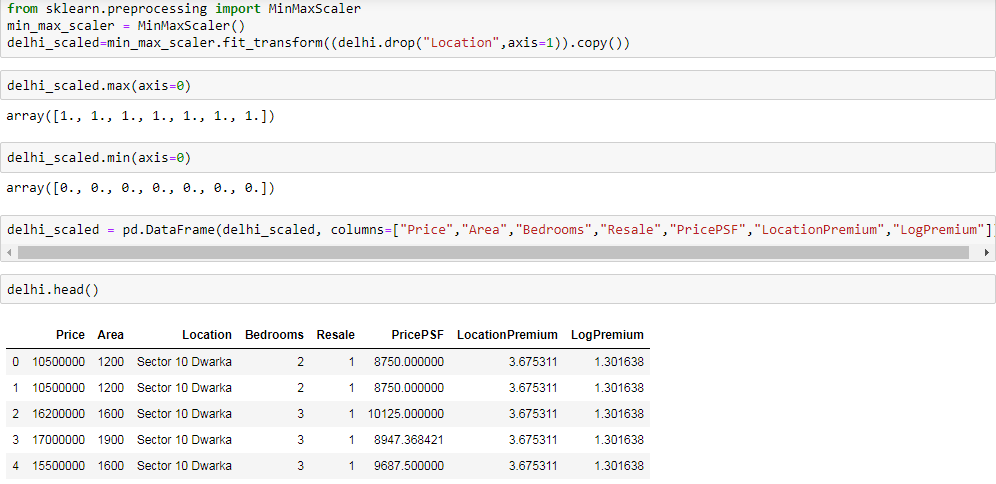
Figure 5.6: Location count

**5.2 Feature engineering:**

Now we will create a pivot table to make location as numerical variable from categorical

 calculating the 'Location Premium' for every Location(PricePSF for every location divided by the minimum PricePSF). This sets the cheapest location as the base location (with a score of 1) and every other location has a premium as a multiple of that base location 

**5.3 Feature Normalization**:



**5.4 Fitting the model:**

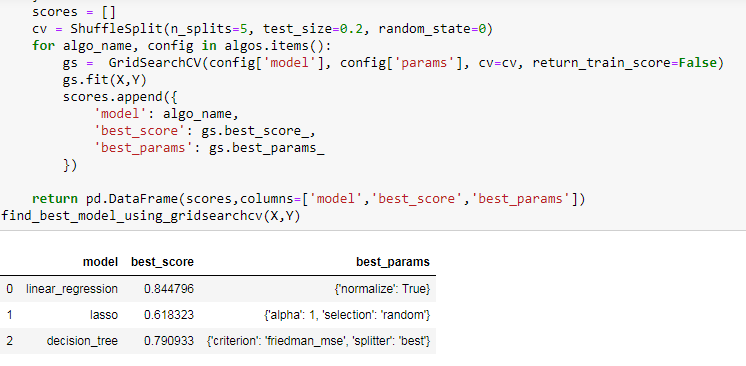
**5.4.1 Separating train and test variables:**



**5.4.2 Grid search**

Grid search to find the best possible parameters of the ML model:





# CHAPTER 6 : RESULTS

Since the accuracy of the model is above 80% we can deploy it and use it to produce precise predictions.

Graph of the locations with similar price range is plotted:

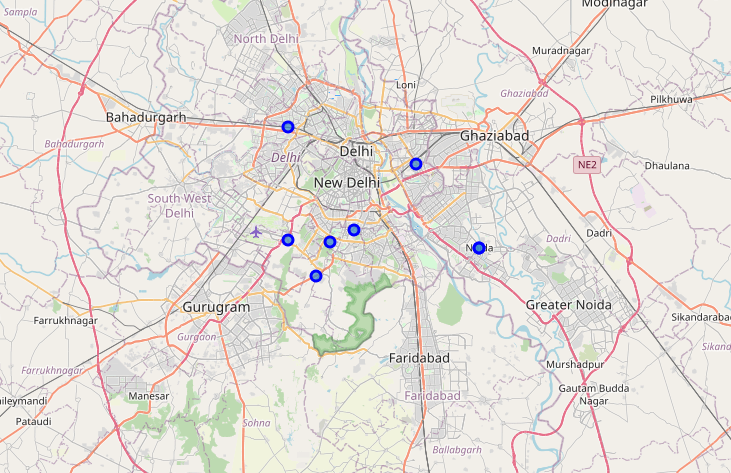


Figure 6.: Similar Price map

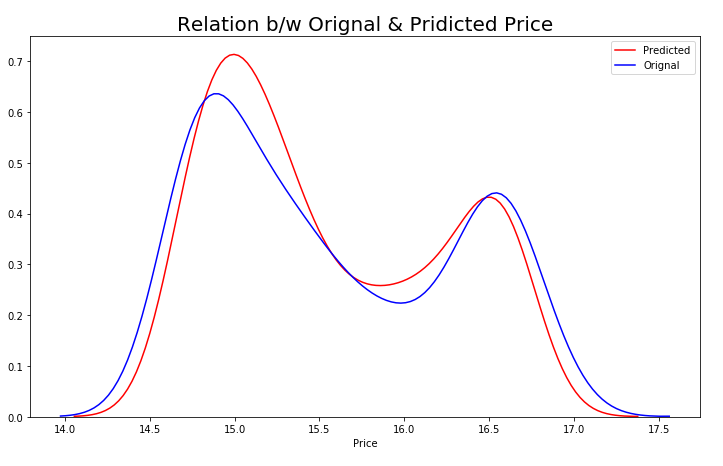


Figure 6.2: Predicted & Original price plot

From the above graph we can conclude that model is a good fit as the predicted and original values almost overlap each other, this also shows that there is no requirement of polynomial regression. So we will proceed with this model.

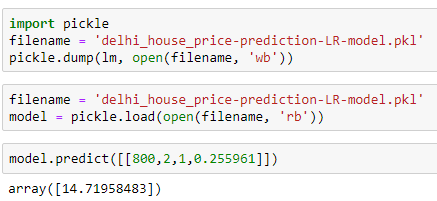
# CHAPTER 7: MODEL DEPLOYMENT

**7.1 Flask implementation**

The model is deployed on a local host it can also be deployed on public host

Picke is used to load the ML mode into the system

Code:

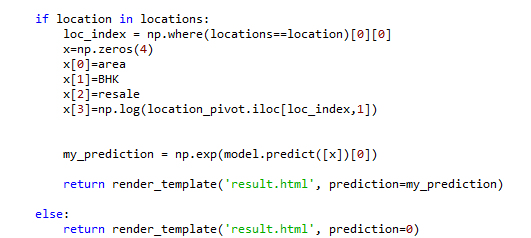


Flask is used to create the web app interface .Django could also be used for this project, but it is advised to not use it as it is a small project



Figure 7.: Flask model deployment

Condition used to make model accept string values:



Other requirements foe the deployment are given bellow

* Flask==1.1.1
* gunicorn==19.9.0
* itsdangerous==1.1.0
* Jinja2==2.10.1
* MarkupSafe==1.1.1
* Werkzeug==0.15.5
* numpy>=1.9.2
* scipy>=0.15.1
* scikit-learn>=0.18
* matplotlib>=1.4.3
* pandas>=0.19

**7.2 Web app interface:**

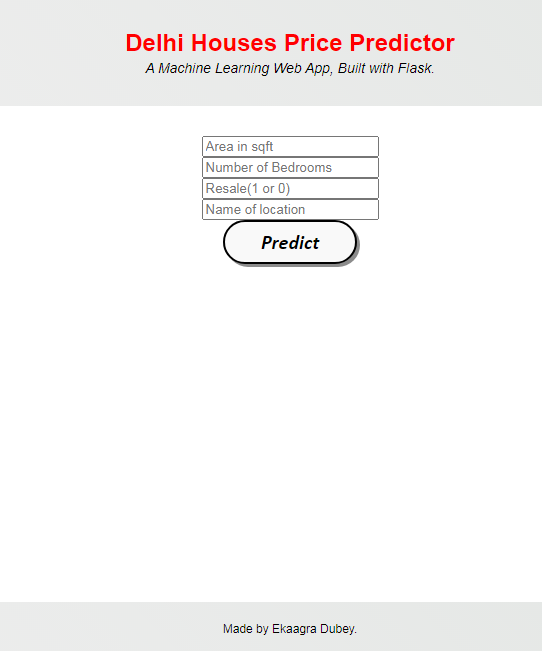


Figure 7.: Web app interface

# 

# CHAPTER 8: CONCLUSION

## 7.1 Conclusion

We have managed to create an end to end machine learning model using best possible model fit that gives users a good approach to predict accurate prices in present and near future

## 7.2 Further Enhancements

* Based on the various parameters and properties files everything from the look and feel to the functionalities can be customized. Thus this project is developed from the beginning with reuse in mind and implicitly uses several design patterns.
* The features provided by use for more interactive enhancement of the screens and inclusion of more data.
* The project can be deployed publicly with more features and better user interface
* Current data set can be concatenated with multiple data and can cover larger area of prediction

# REFERENCE

1. For data and ml algos- https://www.kaggle.com/
2. For python and pandas- https://www.tutorialspoint.com/python\_pandas/index.htm
3. HTML -www.tutorialspoint.com , https://www.geeksforgeeks.org/
4. Flask -www.tutorialspoint.com , https://www.geeksforgeeks.org/
5. Error handling - <https://stackoverflow.com/> , <https://github.com/>
6. CSS -<https://developer.mozilla.org/>