**ASSIGNMENT REPORT [CA-1]**

**(*Project : House Price Prediction and EDA*)**

Submitted by

**(Mogalraj Kushal Dath)**

**Registration Number : 12100559**

**Course Name : Python for Machine Learning**

**Course Code : INT522**

**Section Number : K21ML**

**Roll Number : RK21MLA01**

**Date of Submission : 20-11-2021**

Submitted to

**(Dr. Dhanpratap Singh Sir)**

**School of Computer Science and**

**Engineering**



**House Price Prediction and Exploratory Data Analysis**

**GitHub Link : https://github.com/Kushal11608202/PRJ\_CA1**

**Abstract :**

Nowadays house prices increment consistently, so there is a requirement to design a framework to understand and predict the house prices. House price prediction can help the developer to predict the selling cost of a house and can assist the client in organizing the ideal opportunity to buy a house. The common and main affecting factors on house price are current condition, area, time/year and location of the house.

Housing price patterns are not only the concern of purchasers and venders, however it likewise shows the current financial circumstance. Thus, it is important to predict housing prices without bias to help both the purchasers and venders settle on their decisions. This project utilizes an open source dataset. ([Dataset Link](https://www.kaggle.com/harlfoxem/housesalesprediction/data))

**ACKNOWLEDGEMENT :**

I would like to express my special thanks of gratitude to my mentor Dr. Dhanpratap Singh Sir as well as other higher authorities along with Lovely Professional University who gave me this golden opportunity to do a wonderful project on the topic ‘House Price Prediction’, which also helped me in doing a lot of research and I came to know about so many new things. So, I am thankful to them and extremely privileged to have got all this to complete my project duly.

**Introduction :**

* 1. **Description of the project :**

Investment is a business activity that most people are interested in this globalization era. There are several objects that are often used for investment, for example, gold, stocks and property. Property investment has increased significantly since 2011, both on demand and property selling. At younger generation will need a house or buy a house in the future. Based on preliminary research conducted, there are two standards of house price which are valid in buying and selling transaction of a house that is house price based on the developer (market selling price) and price based on value of selling tax object.

The fundamental problem for a developer is to determine the selling price of a house. In determining the price of a house, the developer must calculate carefully and determine the appropriate method because property prices always increase continuously and almost never fall in the long term or short.

There are several approaches that can be used to do Exploratory data analysis in KC house dataset , one of it is matplotlib and the other is seaborn which are used for the data visualization or data representation in graphical form. And for Prediction of house prices , there is a basic approach called linear Regression will be utilized. To improve or boost the performance we are going to use gradient boosting regression in this project.

* 1. **Limitations :**

There is no guarantee that the data will be contains the exact list of features which affect the prediction of house price. Thus, there might be a risk if the project will be accomplished based only on the public dataset.

Moreover, this project will not cover all regression algorithms; instead, it is focused on the EDA and chosen algorithm, starting from the basic regression techniques (Linear Regression) to the advanced ones (Gradient Boosting Regression).

* 1. **System Design :**

**Diagram

Description automatically generated**

**Libraries :**

1. **Numpy :**

NumPy (Numerical Python) is a linear algebra library in Python. It is very useful for performing mathematical and logical operations on Arrays. It provides an abundance of useful features for operations on n-arrays and matrices in Python. It is the fundamental package for scientific computing with Python. As the whole project is based on whole complex stats ,we will use these fast calculations and provide results.

1. **Pandas :**

Pandas is the most popular python library that is used for data analysis. We will provide highly optimized performance with back-end source code with the use of Pandas.

1. **Matplotlib :**

Matplotlib tries to make easy things easy and hard things possible. We will generate plots, histograms, scatterplots, etc… to make our project more appealing and easier to understand.

1. **Seaborn :**

We will use it for statistical data visualization as Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

1. **Scikit-learn :**

It is a Python library is associated with NumPy and SciPy. It is considered as one of the best libraries for working with complex data. There are a lot of changes being made in this library. We will use it for cross validation feature, providing the ability to use more than one metric. Lots of training methods like logistics regression will be used to provide some little improvements.

**Screenshots and some content based on project :**

Screenshot 1 :

**Text

Description automatically generated with medium confidence**

The screenshot-1 shows about the aim of the project and dataset.

Screenshot 2 :

Table

Description automatically generated with medium confidence

Screenshot-2 shows about the importing the basic required libraries and importing the .csv format dataset to the object df as data frame.

It also presents the first five rows of our dataset using head() function.

Screenshot 3 :

Graphical user interface, text, application

Description automatically generated

Screenshot-3 shows about shape and size of our KC house dataset and some attributes information.

Screenshot 4 :

Graphical user interface, application, table

Description automatically generated

Screenshot-4 shows about the statistical description of the dataset.

Screenshot 5 :

Graphical user interface, text, application

Description automatically generated

Screenshot-5 shows about the information of the data attributes and luckily we don’t have any null values.

Screenshot 6 :

Graphical user interface, text, application, email

Description automatically generated

Screenshot-6 shows about the sum of null values of the dataset attributes in which this case we don’t have any null values so the sum obviously going to be zero.

Screenshot 7 :

A picture containing table

Description automatically generated

Screenshot-7 shows that we are dropping the unnecessary attributes of the data i.e., removing the attributes which does not affect much on house price.

Screenshot 8 :

Graphical user interface, text, application

Description automatically generated

Screenshot-8 shows that the starting point of Exploratory data analysis and some user-defined functions used in our project.

Screenshot 9 :

Chart, scatter chart

Description automatically generated

The plot that we used above is called scatter plot , scatter plot helps us to see how our data points are scattered and are usually used for two variables. The figure tells us about the location of the houses in terms of longitude and it gives us quite an interesting observation that -122.2 to -122.4 sells houses at much higher amount.

Screenshot 10 :

Chart, scatter chart

Description automatically generated

In Scr-10, The figure tells us about the location of the houses in terms of latitude and it gives us quite an interesting observation that 47.6 to 47.7 sells houses at much higher amount.

Let’s see which is most common bedroom number. You may wonder why is it important ? Let’s look at this problem from a builder’s perspective, sometimes it’s important for a builder to see which is the highest selling house type which enables the builder to make house based on that. Here in India , for a good locality a builder opts to make houses which are more than 3 bedrooms which attracts the higher middle class and upper-class section of the society.

Screenshot 11 :

Chart, scatter chart

Description automatically generated

As we can see from the visualization 3 bedroom houses are most commonly sold followed by 4 bedroom. So how is it useful ? For a builder having this data , He can make a new building with more 3 and 4 bedroom’s to attract more buyers.

Screenshot 12 :

Chart, scatter chart

Description automatically generated

Even location influencing the prices of the house. As we can see many houses are sold in between the zip code of 98100 and 96125.

Screenshot 13 :

Graphical user interface, text, application

Description automatically generated

The above graph is a regression plot which gives us graph based on probability density function which is bounded in a contiguous curve.

Screenshot 14 :

Chart

Description automatically generated

The above graph shows a count plot based on number of floors . As we can observe many I floor houses are sold compare to other.

**We can see more factors affecting the price :**

Screenshot 15 :

Graphical user interface

Description automatically generated with low confidence

Screenshot 16 :

Chart

Description automatically generated

Screenshot 17 :

Chart, scatter chart

Description automatically generated

From the above figure we can see that more the Area , more the price though data is concentrated towards a particular price zone , but from the figure we can see that the data points seem to be in linear direction. Thanks to scatter plot we can also see some irregularities that the house with the highest square feet was sold for very less , maybe there is another factor or probably the data must be wrong.

Screenshot 18 :

Graphical user interface

Description automatically generated

The above figure shows about the correlation matrix os the essential factors on house price using heatmap.

Screenshot 19 :

Graphical user interface, text, application

Description automatically generated

Taking Training and Testing data for train test split.

Screenshot 20 :

Graphical user interface, text, application, email

Description automatically generated

**Linear Regression :-**

In easy words a model in statistics which helps us predicts the future based upon past relationship of variables. So when you see your scatter plot being having data points placed linearly you know regression can help you!

This Regression works on the line equation , y=mx+c , trend line is set through the data points to predict the outcome.

The variable we are predicting is called the criterion variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called Simple Regression. and if multiple predictor variable are present then multiple regression.

Let’s look at the code ,

Screenshot 21 :

Graphical user interface, text, application, email

Description automatically generated

So what did we do ? Let’s go step by step :

1. We import our dependencies , for linear regression we use sklearn (built in python library) and import linear regression from it.
2. We then initialize Linear Regression to a variable reg. Now we know that prices are to be predicted , hence we set labels (output) as price columns and we also convert dates to 1’s and 0’s so that it doesn’t influence our data much . We use 0 for houses which are new that is built after 2014.
3. We again import another dependency to split our data into train and test. I’ve made my train data as 80% and 20% of the data to be my test data , and randomized the splitting of data by using random\_state.
4. So now , we have train data , test data and labels for both let us fit our train and test data into linear regression model.
5. After fitting our data to the model we can check the score of our data ie , prediction. in this case the prediction is 43%

Screenshot 22 :

Graphical user interface, text, application, email

Description automatically generated

**Gradient Boosting Regression :**

For building a prediction model , many experts use gradient boosting regression , so what is gradient boosting? It is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

1. We first import the library from sklearn
2. We create a variable where we define our gradient boosting regressor and set parameters to it , here

n\_estimator 🡪 The number of boosting stages to perform. We should not set it too high which would overfit our model.

max\_depth 🡪 The depth of the tree node

learning\_rate 🡪 Rate of learning the data.

loss 🡪 loss function to be optimized. ‘ls’ refers to least squares regression

minimum sample split 🡪 Number of sample to be split for learning the data

1. We then fit our training data into the gradient boosting model and check for accuracy
2. We got an accuracy of 83.85% which is amazing!!!

Screenshot 23 :

Graphical user interface, text, application, email

Description automatically generated

**OUTPUT of the Project :**

Screenshot 24 :

Graphical user interface, chart

Description automatically generated

Screenshot 25 :

Graphical user interface, chart, application

Description automatically generated

**Comparing the Score of Linear Regression and GBR :**

Screenshot 26 :

Chart, bar chart

Description automatically generated

Screenshot 27 :

Graphical user interface, text, application

Description automatically generated

Screenshot 28 :

Graphical user interface, text, application

Description automatically generated

Some Error metrics and Storing the predicted output to Output.csv file.

**References :**

<https://ieeexplore.ieee.org/abstract/document/8473231>

<http://www.mecs-press.net/ijieeb/ijieeb-v12-n2/IJIEEB-V12-N2-3.pdf>

<https://www.tandfonline.com/doi/abs/10.1080/09599916.2020.1832558>

<https://ieeexplore.ieee.org/abstract/document/8882834>