PROJECT REPORT

HOUSE PRICE PREDICTION USING MACHINE LEARNING

ABSTRACT

House/Home are a basic necessity for a person and their prices vary from location to location based on the facilities available like parking space, locality, etc. The house pricing is a point that worries a ton ofresidents whether rich or white collar class as one can never judge or gauge the valuing of a house basedon area or offices accessible. Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation.

This research provides an overview about how to predict house costs utilizing different regression methods with the assistance of python libraries. The proposed technique considered the more refined aspects used for the calculation of house price and provide the more accurate prediction. It also provides a brief about various graphical and numerical techniques which will be required to predict the price of a house. This research contains what and how the house pricing model works with the help of machine learning and which dataset is used in our proposed model.

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

Buying of a house is one of the greatest and significant choice of a family as it expends the entirety of their investment funds and now and again covers them under loans. It is the difficult task to predict the accurate values of house pricing. Our research would make it possible to predict the exact prices of houses. This research is proposed to predict house prices and to get better and accurate results. The stacking algorithm is applied on various regression algorithms to see which algorithm has the most accurate and precise results. This would be of great help to the people because the house pricing ids a topic that concerns a lot of citizens whether rich or middle class as one can never judge or estimate the pricing of a house on the basis of locality or facilities available. To accomplish this task, the python programming language is used. Python is a high level programming language for general purpose programming.

For our research, we have considered Pune as our primary location and are predicting real-time house prices for various localities in and around Pune. In metropolitan city like Pune, the prospective home buyer considers several factors such as location, size of the land, proximity to parks, schools, hospitals, power generation facilities, and most importantly the house price. We have taken into account a verifieddataset with diversity so as give accurate results for all conditions. Regression techniques are widely used to build a model based on several factors to predict price. In this study, we have made an attempt to build house price prediction regression model for data set that remains accessible to the public. We have considered prediction models, they are ordinary least squares model.

1.1. PROBLEM STATEMENT

Given a dataset containing information about houses (e.g., number of bedrooms, square footage, location, etc.) and their corresponding sale prices, the task is to build a machine learning model that can accurately predict the sale price of a new house given its features.

The objective of this problem statement is to create a model that can help potential buyers and sellers make informed decisions about the price of a house. Additionally, real estate agents and property developers can also use this model to estimate the price of a property they are interested in buying or selling.

The model's performance will be evaluated based on metrics such as root mean squared error (RMSE) and mean absolute error (MAE), with the goal of minimizing these metrics and producing the most accurate predictions possible.

1.2. OBJECTIVES:

- To predict the efficient house pricing for customers with respect to their budgets and priorities.
- To develop a model which predicts the property cost for a customer according to their interest.
- To minimize the difference between predicted and actual rating (RMSE/MSE).

CHAPTER 2: THEORETICAL BACKGROUND

2.1. METHODOLOGY:

Detailed analysis of this data set composed of data collection, data cleaning, data visualisation and datapre-processing so that we get a proper data set to work upon. Data collection is the process of gatheringinformation on variables in a systematic manner. We found this dataset on Kaggle, which would suite our project objective. Data Visualization is the graphical representation of information. Data pre- processing is the process of transforming data before feeding it into the algorithm. It is an information mining strategy that includes moving crude information into a justifiable organization. The result of data pre-processing is the last dataset utilized for preparing and testing reason. Data cleaning is the process of detecting and removing errors to increase the value of data.

Linear Regression:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable.

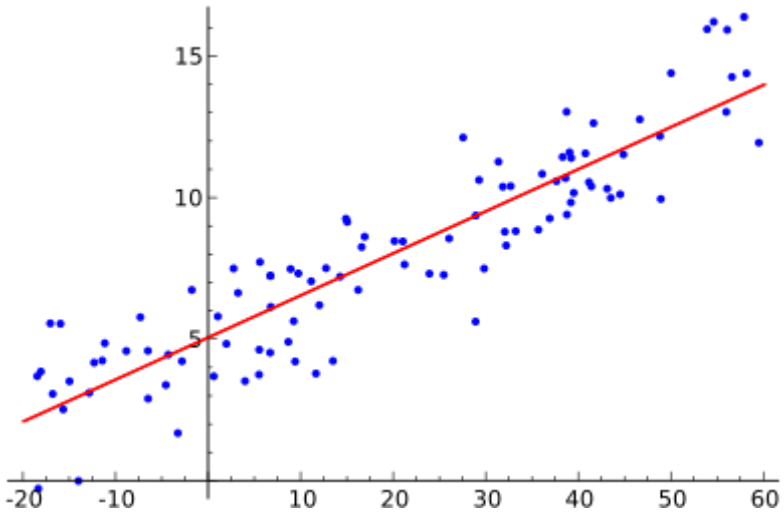

Fig 2.1: LINEAR REGRESSION

Random Forest Regression:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votesof predictions, and it predicts the final output.

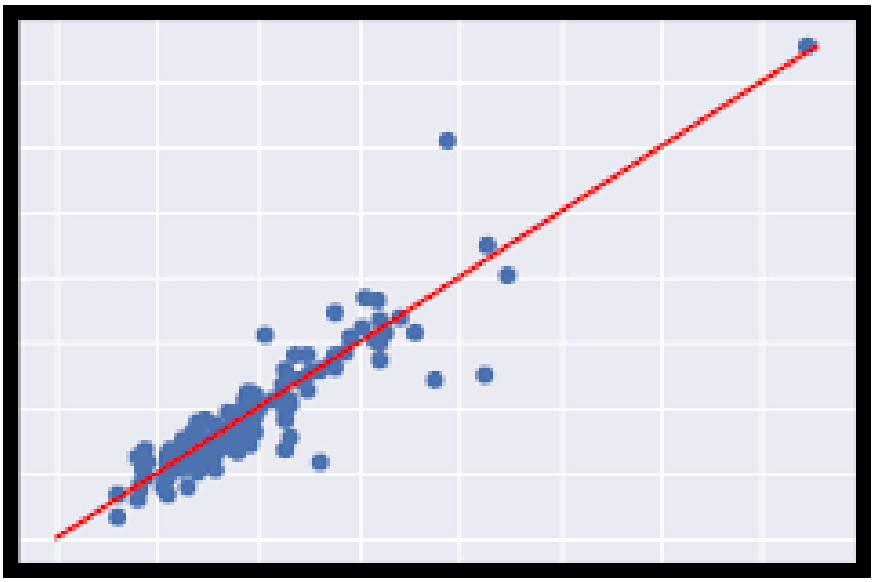

Fig.2.2: Random Forest Regression

CHAPTER 3: EXPERIMENTAL WORK

PROGRAMMING LANGUAGE:

Python is an interpreted, interactive, object-oriented programming language. It incorporates modules, exceptions, dynamic typing, very high level dynamic data types, and classes. It supports multiple programming paradigms beyond object-oriented programming, such as procedural and functional programming.

Python is a computer programming language often used to build websites and software, automate tasks, and conduct data analysis. Python is a general-purpose language, meaning it can be used to create a variety of different programs and isn't specialized for any specific problems.

LIBRARIES:

- Pandas
- Seaborn
- Matplotlib

CHAPTER 4: RESULT

4.1. CORRELATION WITH HEATMAP:

A correlation heatmap is a heatmap that shows a 2D correlation matrix between twodiscrete dimensions, using colored cells to represent data from usually a monochromatic scale. The values of the first dimension appear as the rows of the table while of the second dimension as a column.

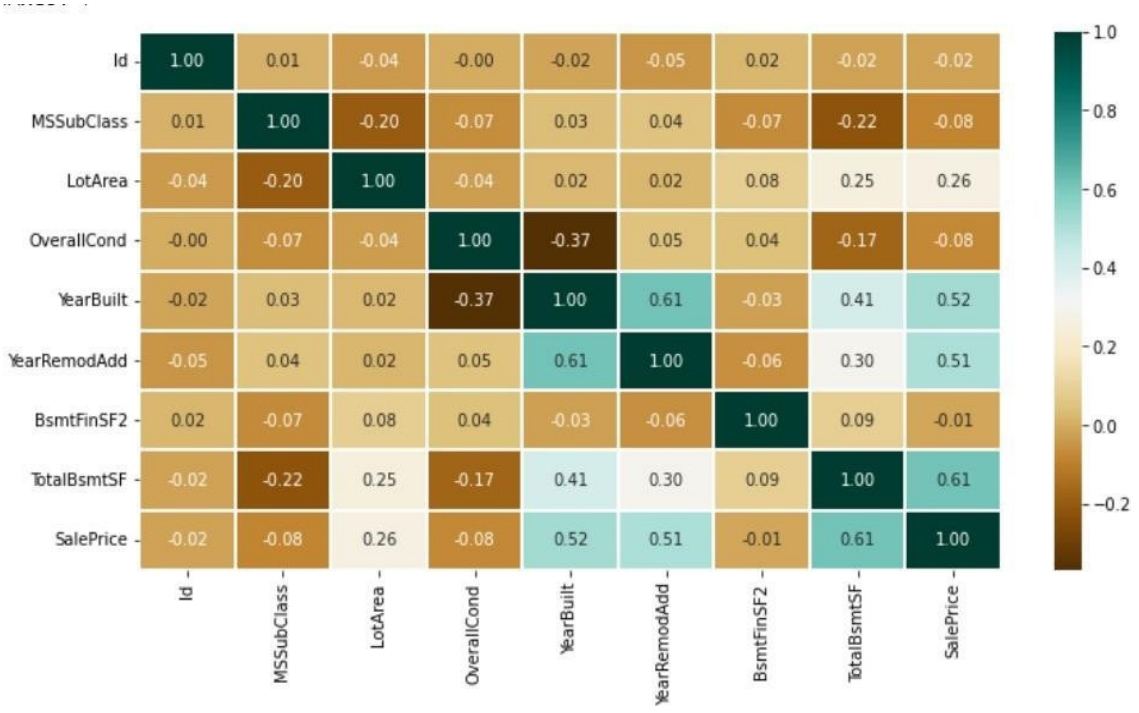

Fig 4.1: Correlation with heatmap

4.2. LINEAR REGRESSION OUTPUT:

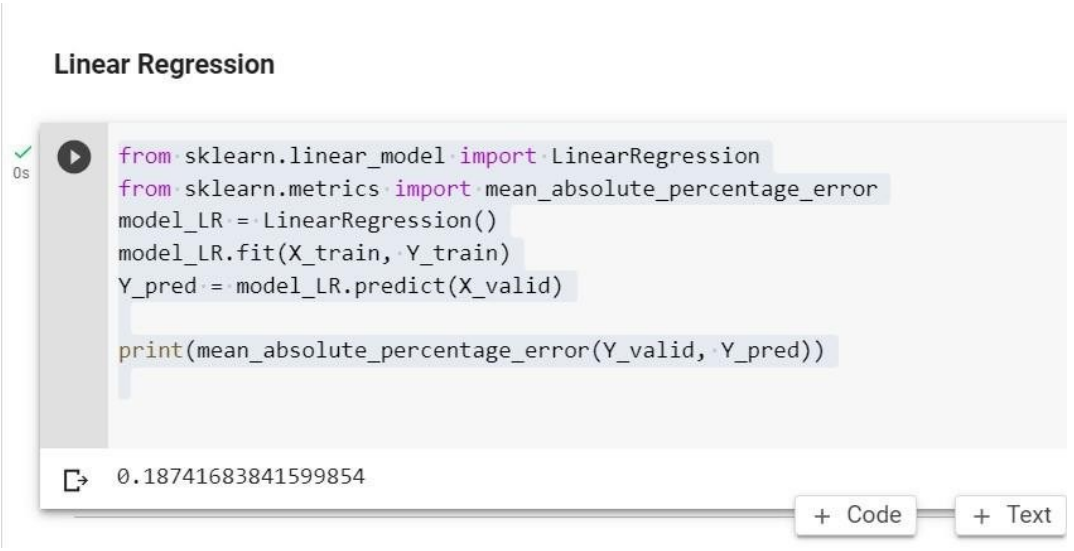

Fig 4.2 LR SCORE

4.3. RANDOM FOREST REGRESSION OUTPUT:

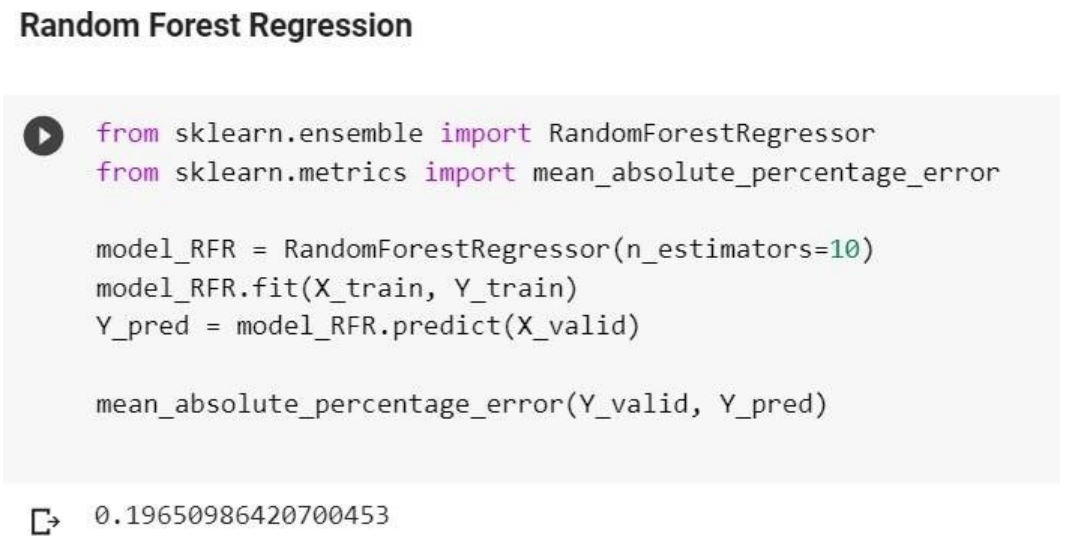

Fig 4.3 RFR SCORE

4.4 MODEL ANALYSIS TABLE:

MODEL NAME	RESULT (in MAPE)R-1	RESULT (in MAPE)R-2
LINEAR REGRESSION	0.18741683841599854	0.18871583863599754
RANDOM FOREST REGRESSION	0.19650986420700453	0.19168874886446152

CHAPTER 5: CONCLUSION

Thus, House Price Prediction data is successfully analyzed using Linear Regression and Random Forest Regression model techniques of Machine Learning.

The sales price for the houses are calculated using different algorithms. The sales prices have been calculated with better accuracy and precision. This would be of great help for the people. To achieve these results, various data mining techniques are utilized in python language. The various factors which affect the house pricing should be considered and work upon them. Machine learning has assisted to complete out task.