Project Synopsis

Name	Maria Wilfred Melvin
USN	222VMBR02629
Elective	Data Science and Analytics
Date of Submission	05.08.2024

Title:

A study on House price prediction in Urban India using machine learning Algorithm

Problem Statement:

Housing is one of the basic components that help to measure the economy of a nation. This determines the success of the country. When there is an increase in the economy, people migrate from urban to rural areas. This resulted in an increase in the population of urban society. An increase in urban society will increase the demand for accommodation. If there is an increase in demand, the price of a house will increase to a great extent. Infrastructural development in the area increases the price of houses. For example, if the area is filled with motorable roads and stable electricity, residential areas get high demand, which leads to an increase in the demand for rental houses in the specific area. Generally, it is quite a complex task to predict house rental prices for investors and valuers. They have to depend on market data to assess the rental prices of the building over a period of time. The outcome has the power to influence the decisions of stakeholders, starting from buyers and investors to sellers. The conventional way of predicting prices is on the basis of historical trends, expert opinions, and comparisons. This fails to capture the dynamic relationship in the real estate market. Thus, the key issue of the study is to make an accurate prediction of house rental that assists investors, prospective owners, developers, appraisers, and others in making valuable decisions.

Objectives:

The objective of the study is

To predict the Indian rental house price by analysing their status of the house

To predict the Indian rental house price by analysing their negotiability

Research Methodology:

The Indian Rental House dataset is available on Kaggle and contains information about rental house price details in India (https://www.kaggle.com/datasets/bhavyadhingra00020/india-rental-house-price/data). This dataset contains more than 13913 data points with 16 variables, including house type, house size, location, city, latitude, longitude, price, currency, number of bathrooms and balconies, which are negotiable, price per square foot, verification date,

description, security deposit, and status. All the variables served as features of the dataset. This helps to predict the price on the basis of status and negotiability.

As soon as the variables are determined, it is necessary to preprocess the data. It is done with three points. One is to check for missing data points. Variables that contain more than 50% of the missing data are removed from the dataset. Get rid of the whole attribute and set the values closer to the mean. Check the outliers of the features; remove the outliers of the features to increase the performance of the model. Next is normalising the numerical values and encoding the categorical values one at a time. The next step is to explore the data and find the appropriate feature using a heat map. The correlation matrix is performed for all the features to find out the most correlated and least correlated variables. The results help to recognize the interaction between the features and target variables. Next is to select the most important features that will be used for the study. When identifying the features, scaling is carried out to ensure the proper selection of the most appropriate features of the study. The lower the features, the lower the performance of the model. The standard scalar function helps to naturally distribute the data within the function. It is now clustering at about 0 with SD1. Next is model development. It represents the core of the rental price process. The data is divided into testing and training. The ratio of training and testing will be 80:20. The training set of 80% inclusion of target variable. During the training phase, the model is fitted to the training set of data, optimised for hyperparameters, and assessed using suitable metrics like mean squared error and R-squared. The prediction models are trained using training sets that provide a wealth of data to show the relationship between features and different objectives. The model gains knowledge that it uses to make predictions. The model is trained with different machine learning algorithms. The results help predict the model with the highest accuracy.

Limitations of the Study:

The data is sourced from the open data source, i.e., Kaggle. This is limited to Bangalore, Pune, and Delhi. The determination of a model is based only on house characteristics. The optimization of the model is based on feature selection, engineering, and tuning methods. The scope of the research is limited to sixteen features. The comparison of model performance is only on the basis of mean absolute error and absolute percentage error. These features may have an explanatory effect on house rental prices. This did not cover all the areas of house rental prices in which ML intervention is possible. This considers the tasks in the rental house price that are suitable for ML techniques. It did not intend to present an exhaustive list of ML

techniques. This study fails to explain how each function performs at ML. This study fails to consider the neighbourhood features, including the seaside, schools, recreation, road network, restaurants, cafes, fire stations, crime rate, and police stations. When considering the features for ML, the accuracy is not determined on the basis of macroeconomic factors including the stock market, balance of trade, producer price index, interest rate, housing starts, and other factors.

Work Plan:

Week No	**Activities Completed**
Week 1	a. Introduction
Week 2	a. Review of literature
Week 3	a. Research gap
Week 4	a. Methodology b. Data preprocessing
Week 5	a. Model development
Week 6	a. Analysis and interpretation
Week 7	a. Findings
Week 8	a. Conclusion and implications