PHASE:3

HOUSE PRICE PREDICTION USING MACHINE LEARNING

REG NO :513421104033

NAME :Rubiga sri S.

Executing the acquisitions
Identification of target market and properties
Building the plan
Evaluate resources

Context:

- The spectrum of housing options in India is incredibly diverse, spanning from the opulent palaces once inhabited by maharajas of yore, to the contemporary high-rise apartment complexes in bustling metropolitan areas, and even to the humble abodes in remote villages, consisting of modest huts. This wide-ranging tapestry of residential choices reflects the significant expansion

- witnessed in India's housing sector, which has paralleled the upward trajectory of income levels in the country. According to the findings of the Human Rights Measurement Initiative, India currently achieves 60.9% of what is theoretically attainable, considering its current income levels, in ensuring the fundamental right to housing for its citizens.
- In the realm of housing arrangements, renting, known interchangeably as hiring or letting, constitutes an agreement wherein compensation is provided for the temporary utilization of a resource, service, or property owned by another party.
- Within this arrangement, a gross lease is one where the tenant is obligated to pay a fixed rental amount, and the landlord assumes responsibility for covering all ongoing property-related expenses.
- The concept of renting also aligns with the principles of the sharing economy, as it fosters the utilization of assets and resources among individuals or entities, promoting efficiency and access to housing solutions for a broad spectrum of individuals.

Content

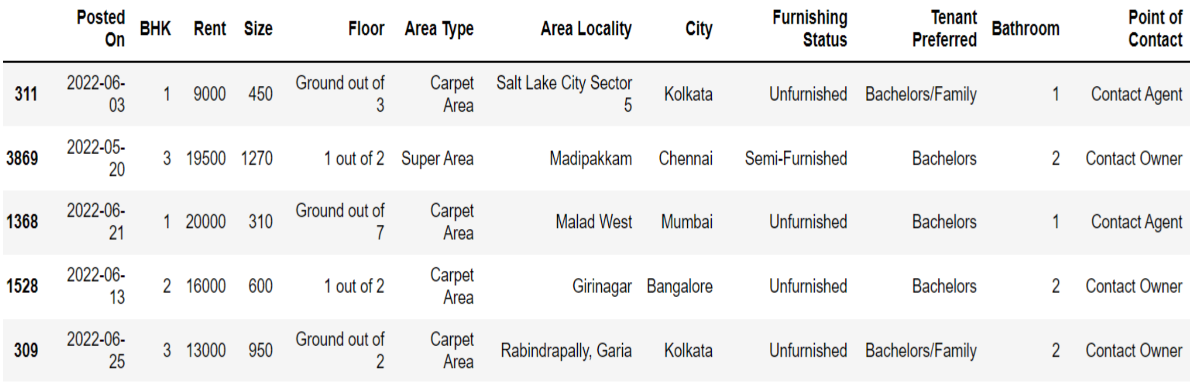
Within this dataset, you will find a comprehensive collection of data pertaining to nearly 4700+ available residential properties, encompassing houses, apartments, and flats offered for rent. This dataset is rich with various attributes, including the number of bedrooms (BHK), rental rates, property size, number of floors, area type, locality, city, furnishing status, tenant preferences, bathroom count, and contact information for the respective point of contact.

Dataset Glossary (Column-Wise)

- **BHK**: Number of Bedrooms, Hall, Kitchen.
- **Rent**: Rent of the Houses/Apartments/Flats.
- **Size**: Size of the Houses/Apartments/Flats in Square Feet.
- **Floor**: Houses/Apartments/Flats situated in which Floor and Total Number of Floors (Example: Ground out of 2, 3 out of 5, etc.)
- **Area Type**: Size of the Houses/Apartments/Flats calculated on either Super Area or Carpet Area or Build Area.
- **Area Locality**: Locality of the Houses/Apartments/Flats.
- **City**: City where the Houses/Apartments/Flats are Located.
- **Point of Contact**: Whom should you contact for more information regarding the Houses/Apartments/Flats.
- **Furnishing Status**: Furnishing Status of the Houses/Apartments/Flats, either it is Furnished or Semi-Furnished or Unfurnished.
- **Tenant Preferred**: Type of Tenant Preferred by the Owner or Agent.
- **Bathroom**: Number of Bathrooms.

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

Structure of the Dataset:



	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
311	2022-06-03	1	9000	450	Ground out of 3	Carpet Area	Salt Lake City Sector 5	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
3869	2022-05-20	3	19500	1270	1 out of 2	Super Area	Madipakkam	Chennai	Semi-Furnished	Bachelors	2	Contact Owner
1368	2022-06-21	1	20000	310	Ground out of 7	Carpet Area	Malad West	Mumbai	Unfurnished	Bachelors	1	Contact Agent
1528	2022-06-13	2	16000	600	1 out of 2	Carpet Area	Girinagar	Bangalore	Unfurnished	Bachelors	2	Contact Owner
309	2022-06-25	3	13000	950	Ground out of 2	Carpet Area	Rabindrapally, Garia	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner

Description:

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them on at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy to enter the market. You are required to build a regression model using regularization in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

Hedonic Pricing:

Hedonic pricing is a price prediction model based on the hedonic price theory, which assumes that the value of a property is the sum of all its attributes value [20]. In the implementation, hedonic pricing can be implemented using regression model. Equation 1 will show the regression model in determining a price. Where, y is the predicted price, and x1, x2, xi are the attributes of a house. While a, b, ... n indicate the correlation coefficients of each variables in the determination of house prices.

DATA SET:

- In this research, we use house price data based on NJOP from Land and Building Tax (PBB) payment structure. Due to limited access to the data, this study used 9 houses data in time series scattered in Malang City area, within 2014-2017.

- Normalization of data is done by completing the empty data at a certain time with the assumption that land prices tend to change every 2 years, while building prices tend to be stable.

- The data tabulation offer information of the houses includes: home id, address (street name), longitude-latitude, year, building area, land area, NJOP building price (IDR/m2), NJOP land price (IDR/m2), distance from city center(km), amount number of campuses, amount number of restaurants, amount number of health facilities, amount number of playground, amount number of schools, amount number of traditional markets or malls, amount number of worship places, and also easiness access to public transportation.

- The city center in this study defined as the location of the square of Malang City. The distance to city center is calculated using Google maps. Meanwhile, easy access to public transportation is calculated between radius 400 meter.

Regression analysis :

The prediction model used in this research is hedonic pricing, the suitable model using regression, with the standard formula as shown in (1). The dependent variable symbolized as Y is NJOP price and independent variables with symbol
x1- x14

consist of year, building area, land area, NJOP land price (IDR/m 2), NJOP building price (IDR/m 2), distance to center of the city, amount number of campuses, amount number of restaurants, amount number of health facilities, amount number of amusement parks, amount number of educational facilities, amount number of traditional markets, amount number of worship places, and easiness to public transportations is shown in (2).

Particle Swarm Optimization (PSO):

PSO is a stochastic optimization method that represents solutions as particle . Amount number of particles are generated randomly, where each particle consists of some dimensions of xi position and velocity vi. Each particle will measure its fitness value..

f (x) = □ from prediction

Where, f (x) is the fitness value of each particle that indicates the error prediction value. Each particle will explore the solution search space to get optimal results. The displacement from one position to another is greatly influenced by the speed of each particle, to obtain the best position required a dynamic speed formulation using

.

$$vi\ t+1 = w.vi\ t + c1\ .\ r1\ (pi – xi) + c2\ .\ r2\ (pgi – xi)$$

Where, vi shows the velocity value for the particle dimension to i to n, t denotes the iteration time, w is the value of the inertia vector whose value is obtained dynamically using (5) [23]. pi is the best position ever obtained for each particle, while the pgi is the best position ever achieved by the whole particle. c1 and c2 sequential are cognitive and social constant, which in this study is 2.5 and 0.5. r1 and r2 are 0.5 and 2.5. Once obtained speed will be updated position..

In the PSO, too fast particle displacement position can make the method fail to obtain the optimum solution. This problem can be handled by performing speed control or velocity clamping [9]. The speed control mechanism by conducting conditions for the speed of each particle uses

. if (vij t + 1 > vj max) then vij t+1 = vj max
if (vij t + 1 < vj min) then vij t+1 = vj min ,

Testing Methods :

The model developed in this research will be tested using several methods such as Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). MAPE is calculated by making an average percentage of the absolute error of each predicted result. Thus, MAPE can indicate how much prediction error. MAPE is described in

Lasso Regression :

Least Absolute Shrinkage and Selection Operator (Lasso) is an L1-norm regularised regression technique that was formulated by Robert Tibshirani in 1996 [6]. 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 a 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 [7].

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 [8]. However, α is determined by using cross-validation.

When $\alpha = 0$,

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 $a \to \infty$ [7]. 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 [7]. Therefore, Lasso is powerful when it comes to feature selection and reducing the overfitting.

Factors affecting prediction:

There are several factors that affect house prices. In his research Rahadi, et al. [14] divide these factors into three main groups, there are

1. Physical condition,
2. Concept and location.
3. Physical conditions are properties possessed by a house that

can be observed by human senses, including the size of the house, the number of bedrooms, the availability of kitchen and garage, the availability of the garden, the area of land and buildings, and the age of the house , while the concept is an idea offered by developers who can attract potential buyers,

For example:

The concept of a

1. Minimalist home
2. Healthy and green environment,
3. Elite environment.

Location:

Location is an important factor in shaping the price of a house. This is because the location determines the prevailing land price [16]. In addition, the location also determines the ease of access to public facilities, such as schools, campus, hospitals and health centers, as well as family recreation facilities such as malls, culinary tours, or even offer a beautiful scenery [17], [18]. In general, the factors affecting the house prices will be presented.

This research aims to create a house price prediction model using regression and PSO to obtain optimal prediction results. PSO is used for selection of affect variables in house prediction, regression is used to determine the optimal coefficient in prediction. In this study, researchers wanted to know the performance of the developed model in time series data. Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location. This research is focused in Malang City, because Malang is one of tourism and urban city in East Java.