House Price Prediction Using Machine Learning

Saloni Raorane[1], Ankita Mishra[2], Devyani Mandwade[3], *Shraddha Dabhade

Department of Computer Engineering
St. John College of Engineering and Management

***Abstract**: In today's world, everyone wishes for a house that suits their lifestyle and provides amenities according to their needs. House prices keep on changing very frequently which proves that house prices are often exaggerated. House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in India to discover useful models for house buyers and sellers. There is a high discrepancy between house prices is most expensive and most affordable suburbs in the city of Mumbai. Moreover, experiments demonstrate that the Multiple Linear Regression that is based on mean squared error measurement is a competitive approach. Output different models showing best score.*

***Keywords:** linear regression, mean squared error*

1. Introduction

Data is the heart of machine learning. Predictive models use data for training which gives somewhat accurate results. Without data we can't train the model. Machine learning involves building these models from data and uses them to predict new data. Machine Learning is a subset of Artificial Intelligence. Machine learning algorithms are purely based on data. Machine Learning algorithms are an advanced version of the regular algorithm. It makes programs "smarter" by allowing them to automatically learn from the data provided by us. The algorithm is mainly divided into two phases and that is the training phase and the testing phase. In Supervised learning, the algorithm consists of a target variable or a dependent variable which is to be predicted from a set of independent variables. Using a function, the inputs are mapped to the desired outputs. The objective of our project is to reduce the problems faced by the customer. In the present situation, the customer visits a real estate agent so that he/she can suggest suitable showplaces for his investments.

2. Related Work

In March 2020[1], Prof. Pradnya Patil, Darshil Shah, Harshad Rajput proposes a method where CatBoost algorithm along Robotic Process Automation is used for real time data extraction. Robotic Process Automation involves use of software robots to automate the tasks of data extraction while machine learning is used to predict house prices with respect to given dataset. In the proposed system first step is data scraping. In this step the structured data can be extracted from the web or any application and saved to a database or any spreadsheet or any CSV file. After data extraction data cleaning is performed. Data cleaning refers to the modifications applied to data before feeding it to the algorithm. Data cleaning is a technique that is used to convert raw data into clean dataset where it deal with missing data, categorical data as per the required needs. After data cleaning algorithms are applied for predicting house rate.

In 2019[2], Parth Ambolkar, Aakash Mane proposes a system in which they have predicted housing price rates using these models i.e. Logistic Regression, Decision Tree and support vector regression with its analogous accuracy and evaluated those based on different types of error metrics like Root Mean Squared Error (RMSE), R-Squared, Mean-Squared Error (MSE), and Mean Absolute Error (MAE). They have used data set which contains 2000 rows with 40 columns which affect house prices. From 40 columns, only 17 were chosen which really affect the price of houses. The area in sq. meters, Location, Year in which house was built, Total BHK, Garage area, swimming pool area, house selling year and selling price of that house. Here dependent variable is S.P. while others are independent variables. Some of the parameters were numeric form and some of them were in a rating form. Rating form was transformed into numeric form. Data processing is the process of transformation of complex data into symmetric knowledge. It finds missing and unnecessary values in the dataset. Whole dataset is searched for Not a Null value and when the row contains null value it will be deleted.

In December 2017[3], Sifei Lu, Xulei Yang proposed a system in which they have used Ridge, Lasso from sklearn and Gradient boosting regression algorithms for house price prediction. The paper also proposes a hybrid Lasso and Gradient boosting regression model to predict individual house price. Ridge and Lasso regressions are used to model cases with large number of features. They have used Kaggle-kc house dataset. Root-Mean-Squared-Error method is used for the measure of the differences between values predicted by a model or an estimator and the values observed. Many iterations of feature engineering have been done to find the optimal no of features. Based on training data, then we tried to find optimal alpha parameter for Ridge regression, optimal alpha parameter for Lasso regression, and a group of optimal parameters for Gradient boosting regression, for example col sample by tree, gamma, max depth, min child weight, subsample, reg alpha, reg lambda, learning rate. Finally, considering

the Coupling effect among regression algorithms, we have evaluated a couple of combinations of hybrid regression to get combination of 65% Lasso and 35 % Gradient boosting.

In 2017[4], Advan Nur Alfiyatin, Hilman Tafiq proposes a system predict house prices based on houses in Malang city with regression analysis and particle swarm optimization (PSO). PSO is used for selection of affect variables and regression analysis is used to determine the optimal coefficient in prediction. The model developed in this research was tested using several methods such as Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

3. Proposed System

First step for starting the coding part is by importing libraries. We have imported pandas, numpy, seaborn and matplotlib library. Every imported library have specific task. Pandas library is basically used for data cleaning and analysis. Numpy library forms the foundation of the machine learning it adds support for large, multi-dimensional arrays and matrices along with a large collection of high-dimensional mathematical functions to operate on these arrays. Seaborn library is used for data visualization and it provides a simple interface and aesthetically better graph plots.

Next step is data cleaning. There can be some error in chosen data set. These errors may negatively impact a predictive model. The process of identifying and correcting errors in the dataset is call data cleaning. It is a good practice to understand the data first and try to gather as many insights from it. Exploratory data analysis is done for performing initial data investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. Every input dataset has some features which are in the form of structured columns. For algorithm to work properly these features should have some specific characteristics. Here, need for feature engineering arises. Feature engineering has two main goals. First one is to make proper and compatible dataset with machine learning algorithm requirements and second one is improving the performance of machine learning models. Sometimes input dataset may contain values which are outside the expected range and unlike other data, these are called outliers. Machine learning modelling can b improved by understanding and even removing these outlier values.

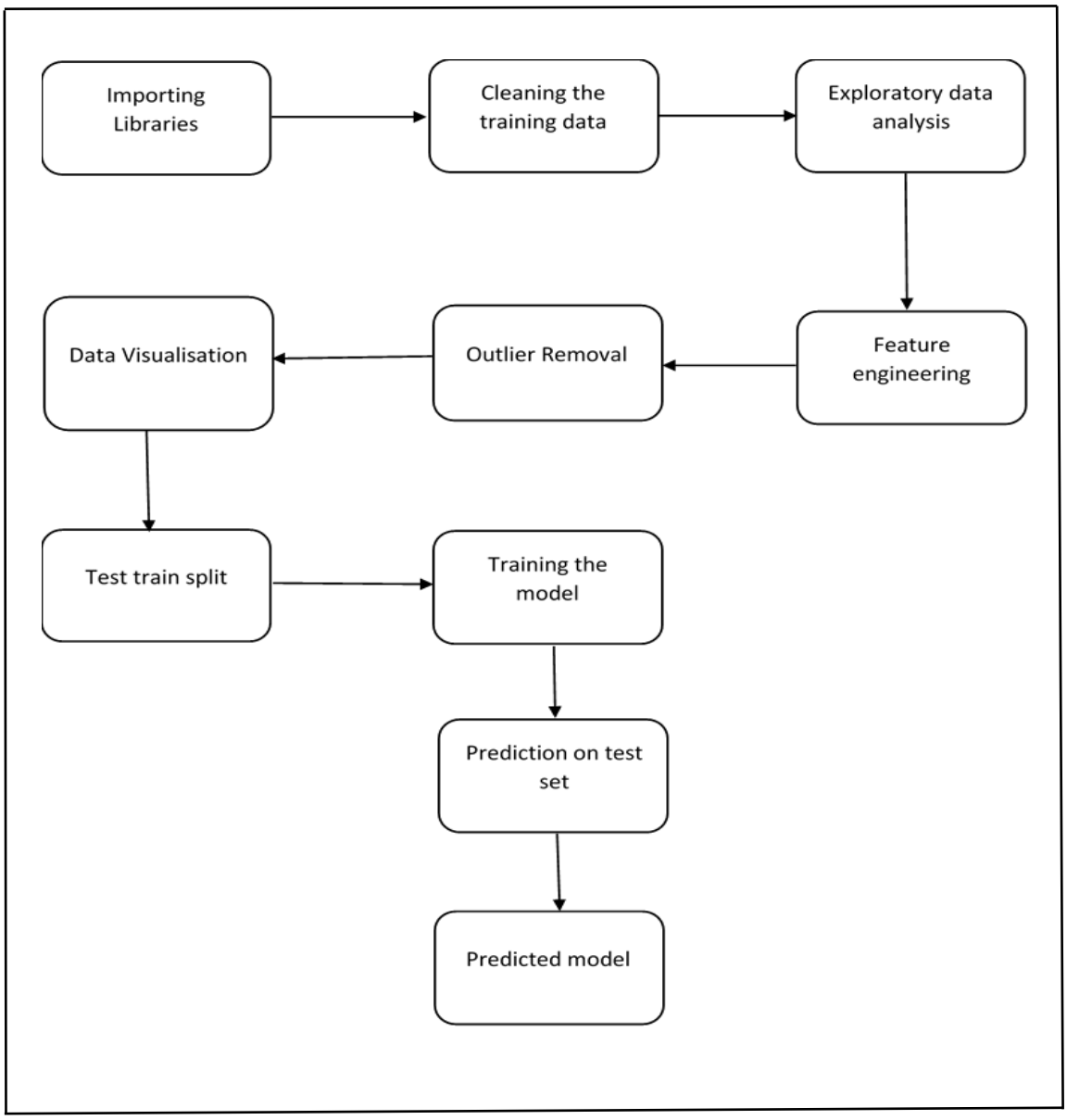

Fig. 3.1 Block diagram for Proposed System

4. Implementation

4.1 Dataset

We are using “Flats for Rent in Mumbai” data set from Kaggle. This data set contains data from 12/08/2019 to 14/01/2020 total of 34349 rows. This dataset contains major key columns like price, latitude, longitude, Locality, City, Description of the house (e.g 3 Bath, Unfurnished,5 floors, East facing 3 bhk flat available for rent in Mira Road East, Mira Road and beyond. It is located in Samarpan, which is a very good society). This data set contains 23 columns but, we will filter the data set and remove the un-relevant column and keep only those columns which will useful to us. we will also be going to apply data cleaning on the filtered dataset. Our Data contains Mumbai Houses only.

4.2 Data Exploration

Data exploration is the process of understanding data with statistical and visualization methods. This step helps in identifying patterns and problems in the dataset, as well as deciding which model or algorithm to use in subsequent steps. Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

```
#understand the data

# Get the no of rows and columns
df.shape
#Get all the column names
df.info()
#getting statistical summary of the Series and DataFrame.
df.describe()
#column labels of the given Dataframe.
df.columns

#Drop features that are not required to build our model**

df1 = df.drop(['city','desc','dev_name','floor_count', 'floor_num', 'id', 'id_string','post_date', 'poster_name',
df1.head()
```

Fig. 4.2 Data Exploration

4.3 Data Cleaning

Data cleaning is the process of identifying the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying, replacing or deleting them according to necessity. Main goal of data cleaning is to identify and remove errors and duplicate data, in order to create a reliable dataset. This process improves the quality of the training data for analytics and enables accurate decision making. Data cleaning is the critical process for the success of machine learning applications. Data cleaning plays a significant part in building a model. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

```
#Data Cleaning:Check for na values
#Verify unique values of each column

df1['user_type'].unique()
df1.shape

#Handling null values
# Get the sum of all na values from dataset
df1.isnull().sum()
df2 = df1.dropna()
df2.isnull().sum()
df2.shape
df2.head()
```

Fig. 4.3 Data Cleaning

4.4 Data Selection

The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses).

4.5 Data Transformation

Data we have may not be in the right format or may require transformations to make it more useful. The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics. The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.

```
#Feature Engineering
df3['price_per_sqft'] = df3['price']/df3['area']
df3.head()
df3_stats = df3['price_per_sqft'].describe()
df3_stats
df3.to_csv("mumbai_home_price.csv",index=False)
df3.head()
df3['bedroom_num'].value_counts()

bathroom_stat = df3['bathroom_num'].value_counts()
bathroom_stat
bathroom_stat_greater_six = bathroom_stat[bathroom_stat<=4]
bathroom_stat_greater_six
```

Fig. 4.4 Data Transformation

4.6 Data Visualization

The representation of data or information in a graph, chart or other visual format is called data visualization. It communicates relationships of the data with images. It is an essential task of data science and knowledge discovery techniques to make data less confusing and more accessible. Visualization takes a huge complex amount of data to represent charts or graphs for quick information to absorb and better understandability. It avoids hesitation on large data sets table to hold audience interest longer. Graphical summary of information makes it easier to identify patterns and trends than looking through thousands of rows on a spreadsheet.

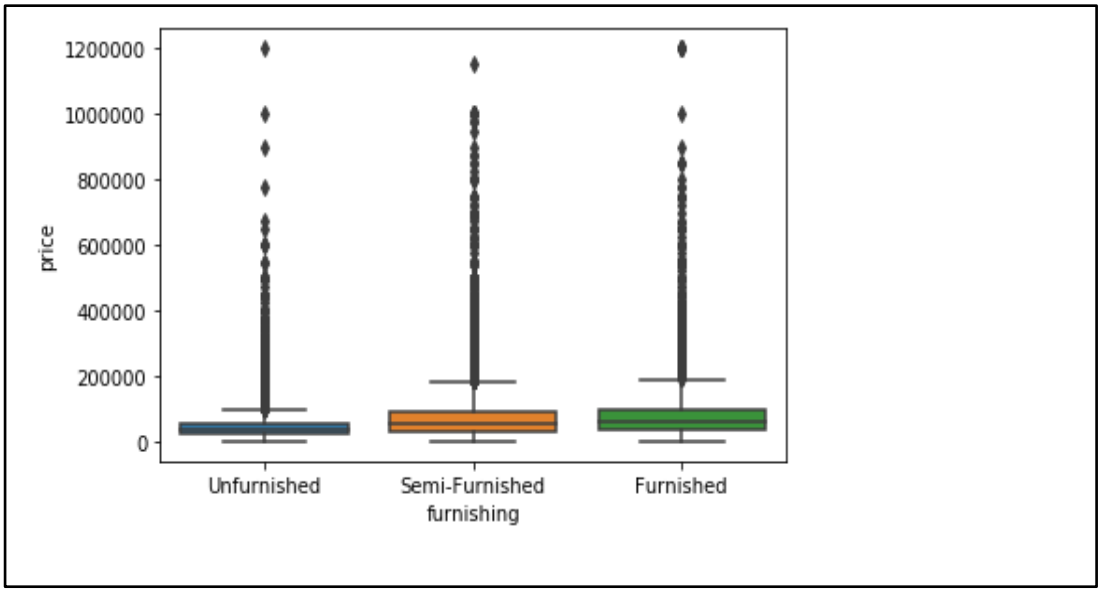

Fig. 4.5.1 SNS boxplot

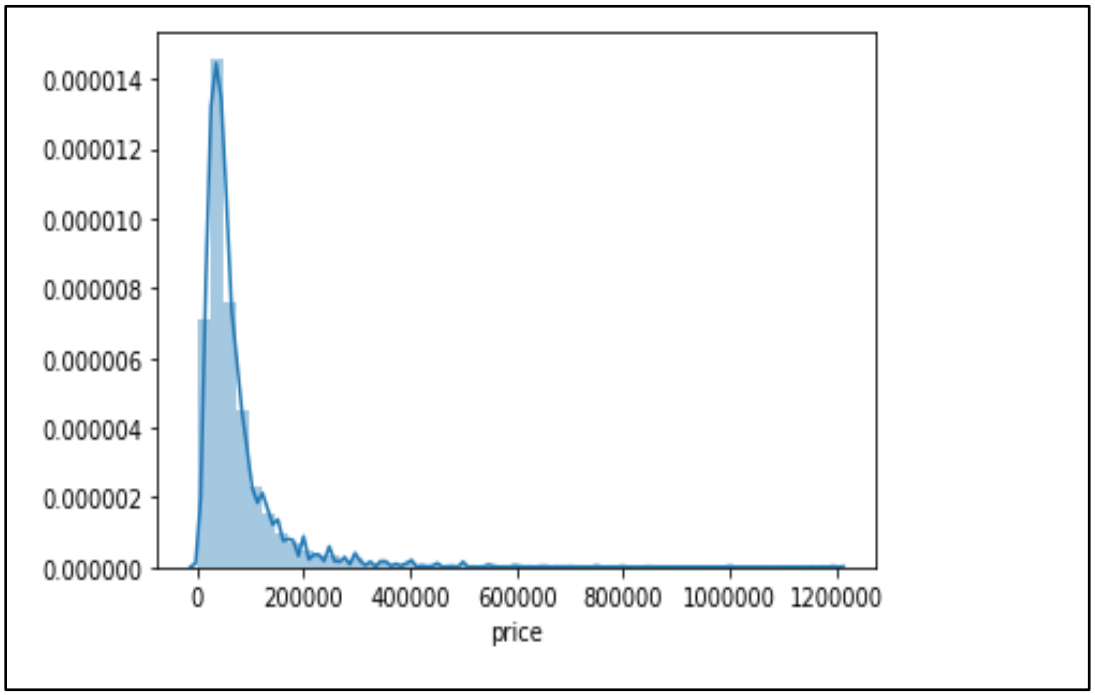

Fig. 4.5.2 SNS Distplot

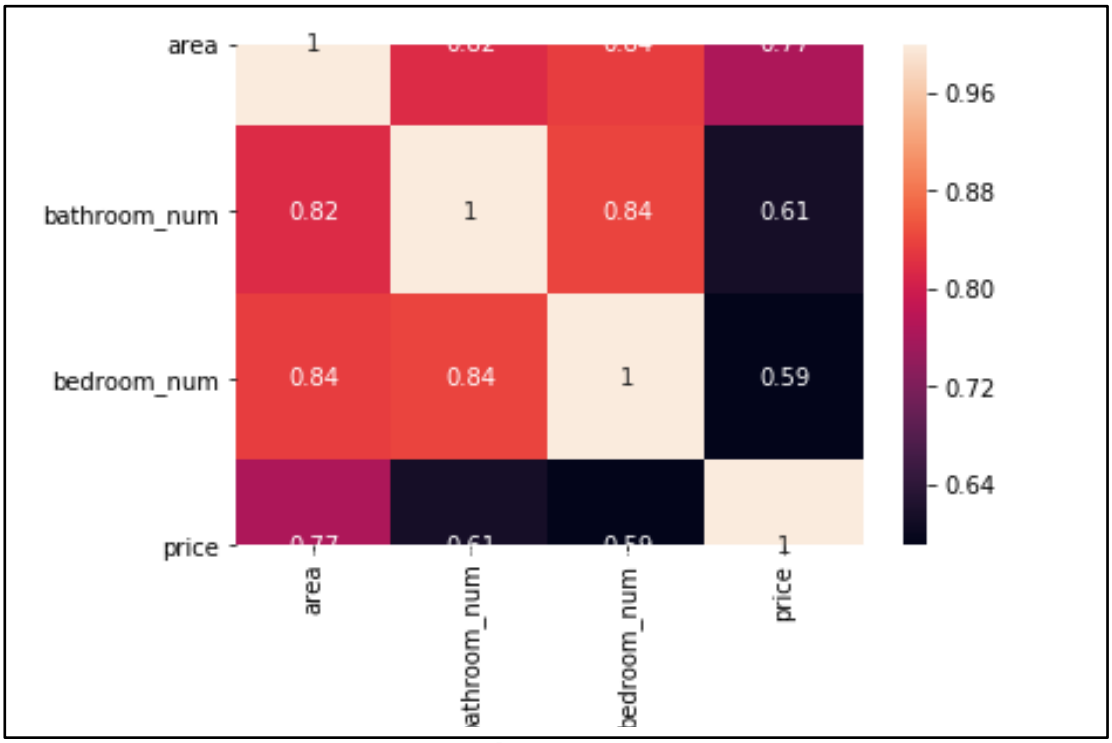

Fig. 4.5.2 Correlation Heatmap

5. Experimental Results

In our project, first we have implemented various machine learning classifiers to know which models is best for our selected dataset. We have calculated the performance of every model and then compared by using dataset. We have used three techniques decision tree, linear regression and lasso regression. Among these three, linear regression is giving the more accurate output which is 0.911502. Whereas lasso regression is giving result as 0.911442 and decision tree giving result as 0.895215.

Model	Best_score	Best_parameters
linear_regression	0.911502	{'normalize': True}
lasso regression	0.911442	{'alpha': 1, 'selection': 'random'}
decision_tree	0.895215	{'criterion': 'mse', 'splitter': 'best'}

5. Conclusion and Future Work

In this approach, house price prediction with help of machine learning is thoroughly described. The proposed approach provides a method to identify which model gives the best and accurate prediction.

After the completion, the system will help customers to find real price of a house in area which they wanted. Prediction of 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. Our initial dataset contains particular area in Mumbai region, after this we can change our dataset so that more and more inside and outside Mumbai region can take help of this system.

6. Acknowledgements

We owe our deepest gratitude and regards towards the ones who offered their valuable guidance in the hour of need. We thank our guide Ms. Shraddha Dabhade for her guidance and precious insights. Her useful comments and feedbacks during the discussions we had and the encouragement to question every technical detail that we came across while completing this project helped us to a great extent.

7. References

[1] *House Price Prediction Using Machine Learning and RPA, Darshil Shah, Harshad Rajput, Jay Cheda, International Research Journal of Engineering & Technology(IRJET),**(2020)***

*[2] House Price Prediction Using Various Machine Learning Algorithms, Parth Ambalkar, International Journal of Advance Research, Ideas and Innovations inTechnology, **(2019)***

*[3] A Hybrid Regression Technique for House Price Prediction, Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, Institute of High Performance Computing (IHPC), **(2017)***

*[4] Housing Prices Prediction with Deep Learning and Random Forest Esemble, Adyan Nur Alfiyatin, Hilman Taufiq, International Journal of Advanced Computer Science and Applications, (IJACSA), **(2017)***