Interactive Visual Exploratory Data Analysis (Eda) For Predicting House Prices In Metro City

Dr. Vidyullata S. Jadhav
Assistant Professor
V. P. Institute of Management Studies & Research, Sangli
vsjadhav@vpimsr.edu.in

Dr. Sampada Gulavani
Associate Professor
Bharati Vidyapeeth (Deemed to be University), Pune
Institute of Management,Kolhapur
sampada.gulavani@bharatividyapeeth.edu

ABSTRACT

House price prediction is becoming increasingly popular, and research teams are increasingly doing similar studies utilizing deep learning or machine learning models. People are quite cautious when it comes to buying a new house based on market techniques and their budgets. Methods for exploratory data analysis can show the structure of data. Due to the abundance of available methodologies, picking ones that will function well together and facilitate data interpretation is a difficult issue. In this paper, we provide a well-fitting collection of tools for conducting a thorough exploratory investigation of a House price prediction dataset. The suggested technique consists of numerous steps, including robust data normalization, outlier handling, feature importance, missing value handling, and so on. By examining several criteria such as the house's area, square footage, number of bedrooms, and so on. The dataset from Kaggle was used in this study. Various graphical display techniques are used to undertake exploratory data analysis. This EDA is critical for using machine learning approaches and constructing models. According to the results of the investigation, the Kharghar location in Mumbai has the most available properties for sale. Also identified are the top ten Mumbai localities with the highest average property price for those who can afford luxury homes in Mumbai. EDA also aids in determining the relationship between the features that are significant in machine learning. Finally, the researcher used the folium library to create an interactive Mumbai map displaying the number of residences for sale by location.

Keywords: House, price, exploratory data analysis, prediction, machine learning, feature importance, folium, interactive

Introduction

Mumbai, which serves as the financial hub of India, is a prime location for investments and attracts end-user interest throughout the entire year. In addition to the employment prospects, Mumbai also provides investment chances in all price ranges; however, potential purchasers at the market for

budget properties should be prepared to look in the city's peripheries. Prices of real estate in Mumbai can fluctuate widely depending on the type of infrastructure, level of connection, level of demand, and level of supply in a certain area. According to the data provided by Makaan.com, the yearly price growth in Mumbai is believed to have increased by 565 counts, and the annual rental yield has also increased by 3.853055% of the previous year. The field of artificial intelligence encompasses a subfield known as machine learning. The machine is learning by itself and testing through the existing dataset using certain methods. As a result, the machine produces results that are more accurate than the manual job would have been. In order to make predictions about new output values based on new input values, machine learning algorithms employ previously collected data as input. The broad field of machine learning can be broken down into two primary categories: supervised learning and unsupervised learning. In the process of supervised learning, a computer programme is educated using an existing database in order for it to be able to make accurate predictions when presented with new information. Unsupervised learning is a type of machine learning in which the computer does not have an instructor present to guide it as it searches for hidden relationships and patterns in the data. The cost of housing has a substantial impact on the economy, and the value ranges it fluctuates within are a major source of concern for both consumers and real estate professionals. The price of housing is increasing on an annual basis, which ultimately strengthened the demand for a method or technique that might estimate the price of housing in the future. There are a variety of elements, such as the physical qualities of the home, its location, the number of bedrooms, and others, that can affect the price of a property. Predictions are traditionally formulated using these elements as their foundation. However, in order to make accurate predictions using such methodologies, one must have sufficient prior knowledge and experience in the relevant field. Techniques from the field of machine learning have been an important contributor to the development of more complex methods for analyzing, forecasting, and visualizing housing values. EDA, or exploratory data analysis, is a method that can summaries the data by identifying the most important aspects of the data and then visually representing that data in the appropriate manner.

Review Of Literature

Recent studies have focused on price prediction performance comparison between hedonic-based methods and machine learning algorithms. (Kauko, 2002) The authors investigated neural network modeling by applying it to the real estate market in Helsinki, which is located in Finland. Their findings suggested that distinct aspects of the development of housing sub-markets could be discovered by sifting through the dataset in search of recurring trends. In addition to this, they presented evidence of the categorization capabilities of two other neural network methods: learning vector quantization while utilizing a self-organizing map. (Fan, 2006) When investigating the connection between house prices and dwelling attributes, authors proposed a number of tree-based methods that serve as useful statistical pattern detection tools. (Liu, 2006) A fuzzy neural network prediction model that is based on hedonic price theory was proposed by the authors as a means of estimating the right price level for newly constructed real estate. The findings of the

experiments suggested that the fuzzy neural network prediction model possessed a powerful function approximation ability and was appropriate for use in the prediction of real estate prices. (Selim, 2009) Both hedonic regression and artificial neural network models were tested for their ability to make accurate predictions, and the results were compared. This research showed that artificial neural network models can be an improved alternative for predicting the price of real estate in Turkey.

In another study, (Aytekin, 2010) The authors proposed a fuzzy logic model as a means of estimating the final selling price of newly constructed homes. (Azadeh, 2012) In order to solve the issue of forecasting and optimizing fluctuations in the housing market, the authors developed a hybrid algorithm that is based on fuzzy linear regression and a fuzzy cognitive map. Over the course of the past few years, a number of researchers have conducted studies on performance comparisons between different machine learning algorithms in an effort to develop more accurate predicting models. (Gerek, 2014) The authors proposed two distinct adaptive neuro-fuzzy (ANFIS) methodologies in order to estimate the selling price of houses in the construction industry. According to the findings of the experiments, the ANFIS version with grid partition models performed significantly better than the ANFIS version with sub clustering models. The ANFIS method with grid partition approach has proven to be effective when applied to the task of estimating housing market values in the building industry. (Wang, 2014) Models for estimating future real estate prices based on particle swarm optimization (PSO) and support vector machines were proposed by the authors (SVM). The results of the experiments showed that the PSO–SVM based real estate price forecasting model suggested by the authors has good forecasting performance when compared to grid and genetic algorithms. (Gu, 2011) The authors suggested a G-SVM technique, which is a combination of genetic algorithms and support vector machines, in order to estimate home prices. According to the findings of the experiments, the predictive accuracy of the G-SVM is significantly higher than that of traditional approaches. However, relatively little work has been done to construct a more accurate model for predicting home prices by comparing the effectiveness of a variety of machine learning algorithms.

Objective Of The Study

The ultimate purpose of the research is to develop an EDA for a house price prediction model that is able to accurately forecast property prices in major metropolitan areas such as Mumbai. EDA places a greater emphasis on evaluating the assumptions that are necessary for the model fitting process, as well as addressing missing values and transforming variables as necessary. EDA provides a short description of the number of rows and columns in the data set, as well as any missing data, data kinds, and a preview of the feature's significance. Bar charts, histograms, and box plots are examples of data visualization techniques used by EDA. Determine and display the associations (correlations) between the variables using tools like heat maps and the like.

DATA AND METHODOLOGY

The Kaggle dataset on House Price prediction for Metro city Mumbai in India is the source of the information that we use. There are 7719 individual data points in all (houses). In addition to the price of the house, we also provide information about the resale price of the house, the maintenance staff, the swimming pool, the jogging track, the rain water harvesting, the indoor games, the shopping mall, the intercom, the ATM, the club house, the school, the 24X7 security, the power backup, the car parking, the staff quarters, the cafeteria, the multipurpose room, the hospital, the washing machine, the gas connection, the air conditioning, the

In order to analyze this dataset, we use EDA. We are going to do the analysis multiple times on various feature sets. Some sets just include the most essential components that are connected to the houses. By way of contrast, we are able to demonstrate how EDA may make use of those data in order to provide estimations that are more accurate overall. The following table displays some of the attributes of the dataset along with their corresponding values.

index	Price	Area	Location	No. of Bedrooms	Resale	Maintenance Staff	Gymnasium	Swimming Pool	Landscaped Gardens	Jogging Track	Rain Water Harvesting	Indoor Games	Shopping Mall	Intercom	Sports Facility	ATM	Club House	School	24X7 Security	Power Backup
0	4850000	720	Kharghar	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1
1	4500000	600	Kharghar	1	1	1	1	1	0	1	1	0	0	0	0	0	1	0	1	1
2	6700000	650	Kharghar	1	1	1	1	1	0	1	1	0	0	1	0	0	1	0	1	1

Mumbai House Price Dataset with first three records

Experiment

The purpose of the experiment is to perform some preliminary processing on the data and assess the level of accuracy achieved by the models. The experiment consists of a few different stages, all of which must be completed in order to obtain the prediction results. The following descriptors can be used for these stages: The datasets will be examined, and then pre-processed, using the procedures that have been established for working with data. Therefore, the preprocessing is done by a series of rounds, with each time an evaluation of the accuracy being carried out using the relevant combination. Splitting the dataset into two portions such that one may be used to train the model and the other can be used for evaluation is referred to as "data splitting." The dataset will be divided such that 80% of it will be used for training and 20% will be used for testing. The Pearson Coefficient Correlation will be used to determine whether the available features have a negative, positive, or zero correlation with the house price. This will determine whether the correlation between the available features and the house price is positive, negative, or neutral.

Exploratory Data Analysis With Visualizations

The financial hub of India, the city that never sleeps, and the city that homes the dreams of millions of people, Mumbai is the city that most exemplifies the concept of a city of dreams as a metaphor. Let's have a look at the top 10 residences in Mumbai that are the most opulent and expensive, as well as the price figures for those homes.

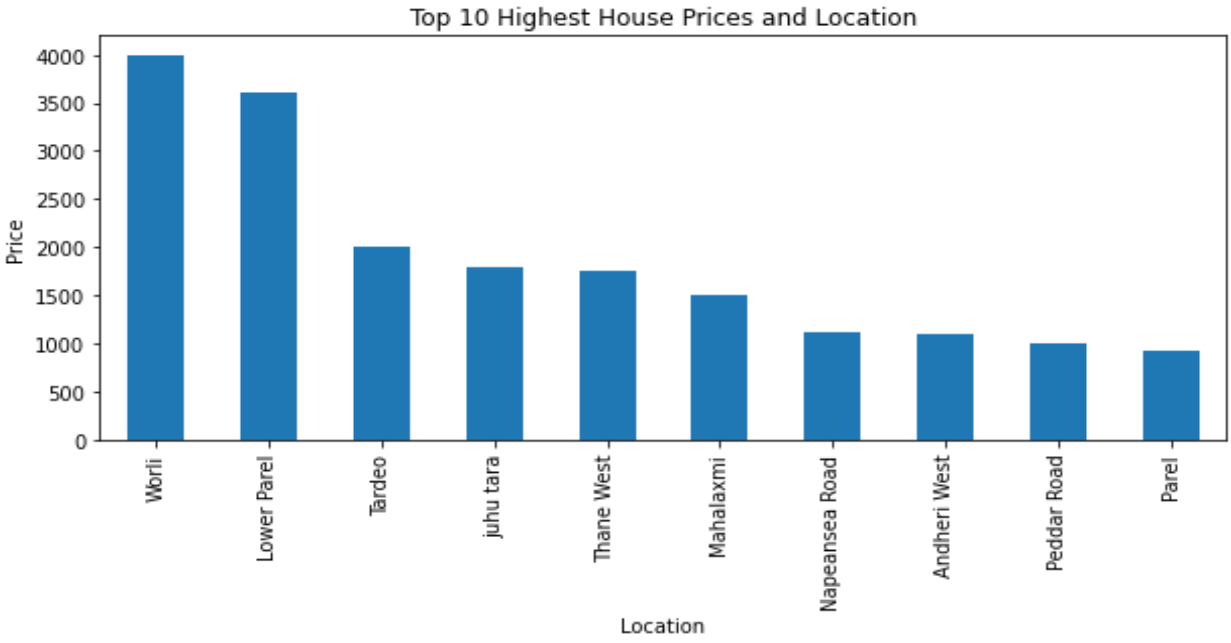


Figure 1: Top 10 Highest House Prices in Mumbai

Worli has the highest average house price in Mumbai, at 4000 lakhs, followed by Lower Parel, which has an average cost of 3600 lakhs, Tardeo, which has 2000 lakhs, Juhu tara, which has 1800 lakhs, Thane West, which has a price of 1750, and Mahalaxmi, which has an average cost of 1500, and then Napeansea Road, which has 1125 lakhs, Andheri West, which has 1100 lakhs, All of these areas have what's known as a live ability quotient, which indicates that social infrastructure such as schools, colleges, hospitals, retail malls, hangouts, provision stores, and other conveniences are within a reasonable distance. These regions have an acceptable amount of physical infrastructure, such as linking roads, travel and transport facilities, which is one of the reasons why they have such a high livability score.

The following scattered plot determines the distribution on house prices in Mumbai with built-up area and number of bedrooms available figure (2).

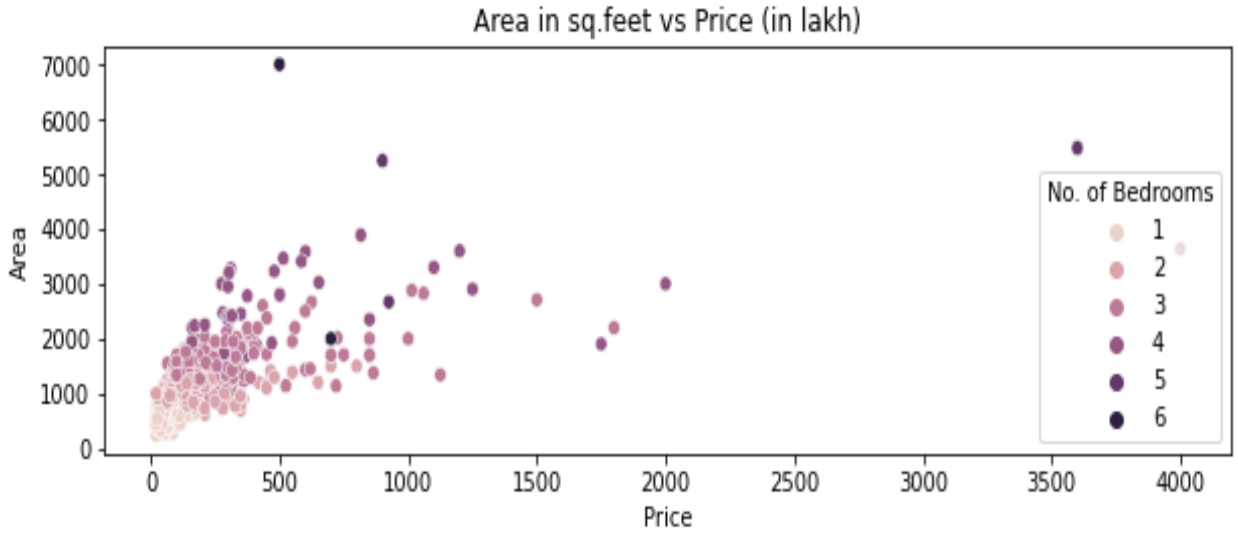


Figure 2: House prices in Mumbai with built-up area and number of bedrooms

The scatter plot of the sale price is presented for our perusal in Figure 2. The majority of the points are put together on the underside. In addition, it would appear that there are no major anomalies in the sale price variable. According to the data presented, it is abundantly evident that the price of a house is mostly determined by the number of bedrooms that are available and the total square footage of the built-up space. The accompanying chart, Figure 3 provides an overview of the total number of real estate options accessible in relation to Mumbai's many localities.

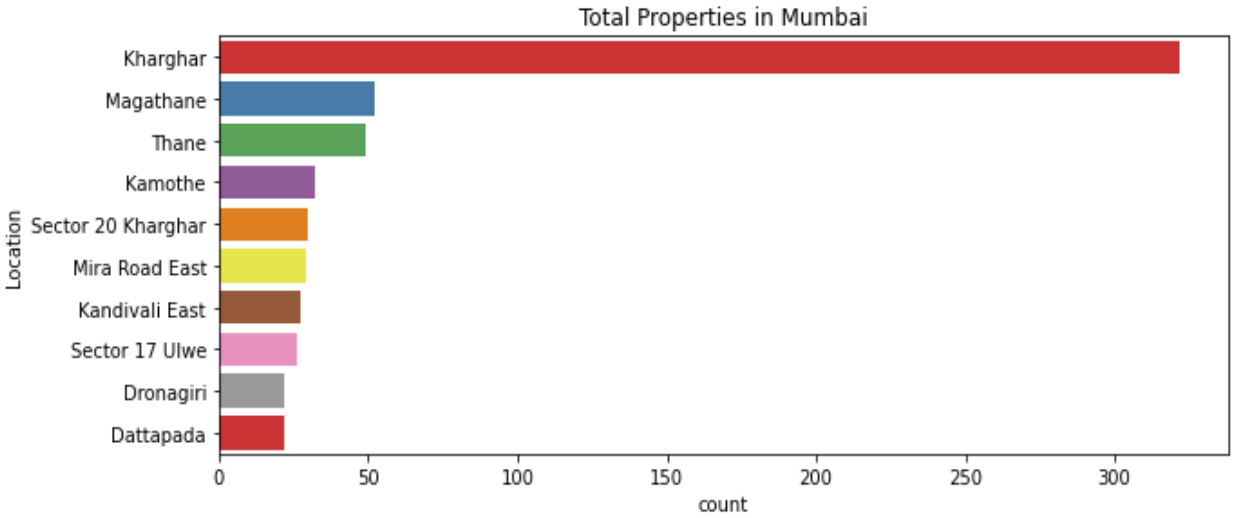


Figure 3: Total number of properties

Based on the statistic, it is possible to infer that the locality of Kharghar in Mumbai contains the greatest number of homes that are now available for purchase, followed by Magathane, Thane, and Kamothe. The researchers had been using ELI5, which is a Python module that is used to analyse machine learning classifiers and explain their predictions. This was done so that the visualisation could be as effective as possible. Debugging algorithms like sklearn regressors and classifiers, XGBoost, CatBoost, and Keras, among others, is a common application for this tool. The chart that follows presents a comparison of the availability of new houses and houses that have been on the market for some time in each of Mumbai's localities. The buyer and seller will be able to locate these properties more quickly and efficiently as a result of this.

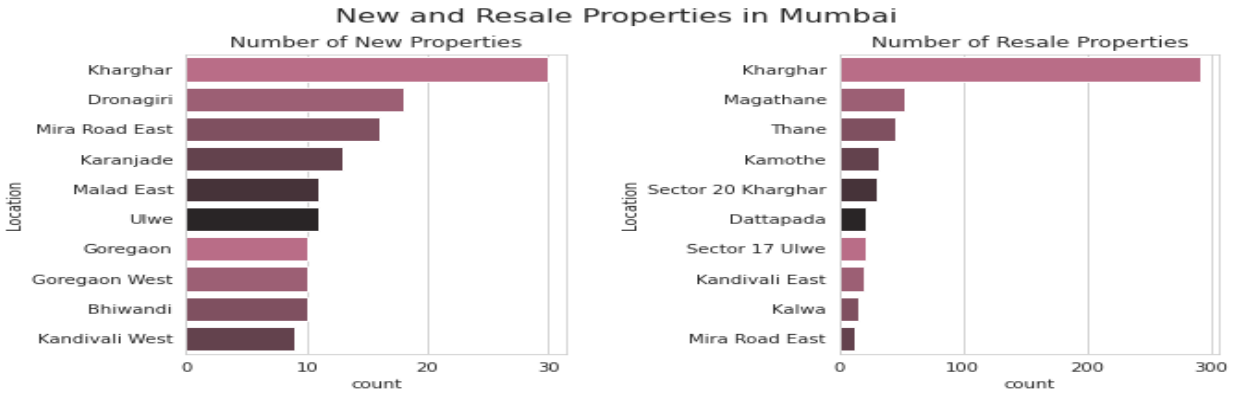
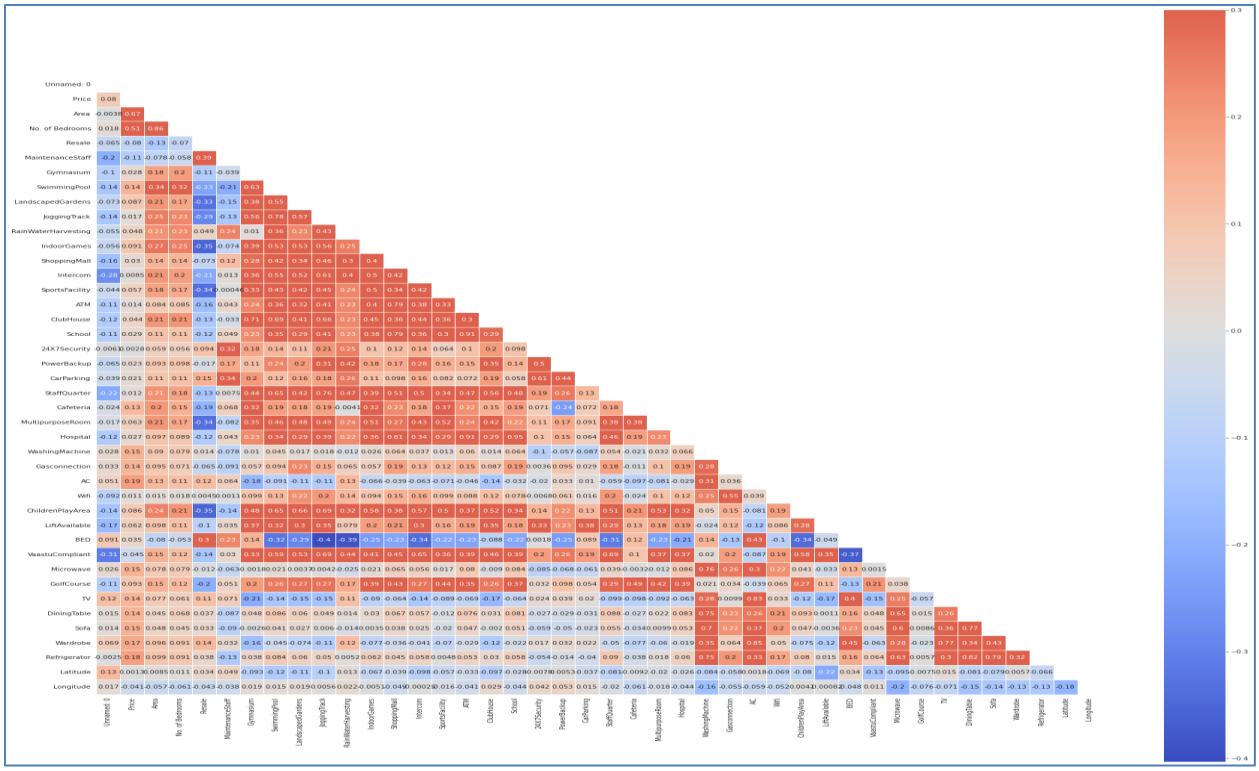


Figure 4: New and Resale Houses in Mumbai

Correlations Plot for the most related numeric predictors, The researchers opted to create a correlation plot in order to determine which numeric variables have a high degree of association with the Price. When modeling with linear regression, it is vital to get rid of variables that are correlated with one another in order to increase the accuracy of the final prediction model. The correlation plot is shown in figure 6.



According to the hue, orange has the highest correlated value, which means that it has a positive link with price. On the other hand, blue indicates that the predictor is not associated at all, which suggests that it has a negative relationship with the price of the house. And it should be obvious that the majority of these elements have an effect on the prices when purchasing a house. And the factor with the highest correlation is area, followed by the number of bedrooms that are available, followed by the proximity of hospitals and schools.

Figure 5 demonstrates that the distribution of sale prices is right skewed, which indicates that the distribution of sale prices isn't normal. This indicates that the distribution of sale prices isn't normal. It makes sense given that very few people have the financial means to purchase extremely pricey homes. Before conducting model fitting, researchers are required to perform transformations on the sale prices variable.

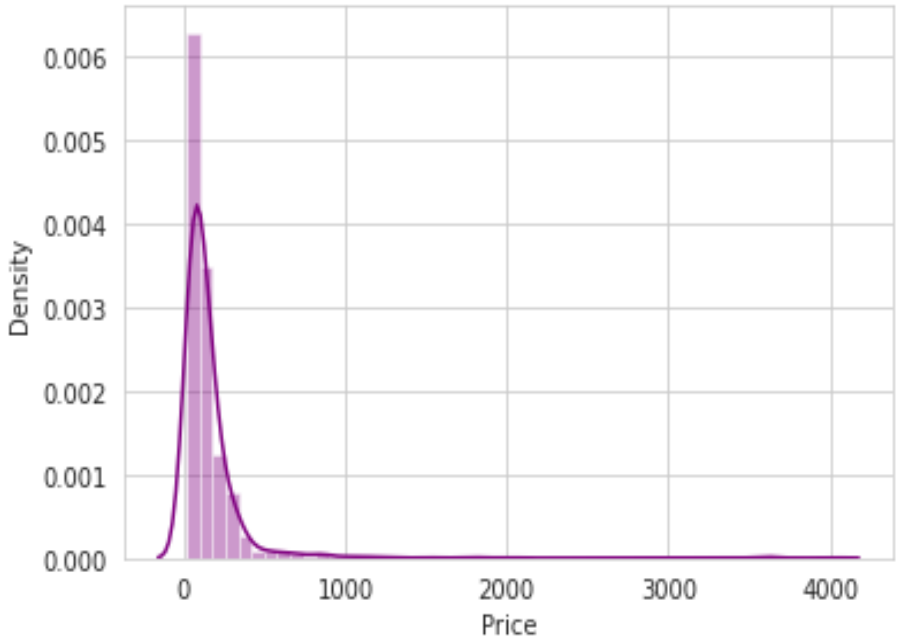


Figure 5: New and Resale Houses in Mumbai

Interactive Visual For Houses For Sale : Mumbai Map

Researchers had referred to the geopy library in order to obtain the latitude and longitude coordinates of the locations of the homes that were on the market in Mumbai. Then, we created the interactive base map of Mumbai using the "folium. map" method that is included in the folium package in order to display the amount of homes that are up for sale in each individual location.

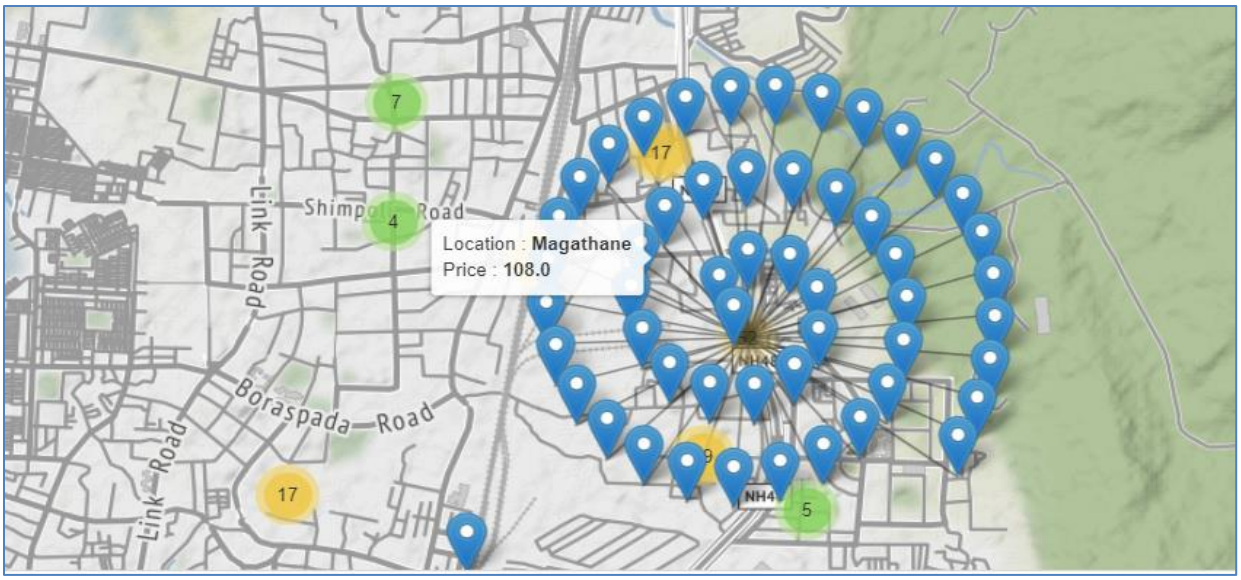


Figure 5: Houses Available for sale with Location and Price in Mumbai

Handling The Outliers

Treating outliers is an essential part of the data preprocessing step, which we are now in. Managing the data's anomalies, also known as extreme values, requires careful attention. Using a Boxplot, the researcher was able to identify the outliers in the area column of the dataset. These data points are considered outliers since they are not included in the box of other observations, which means they are nowhere near the quartiles. The figure below displays data points ranging from 2000 to 8000. In this case, we performed a Uni-variate outlier analysis, which means that we checked for anomalies based just on the Area column.

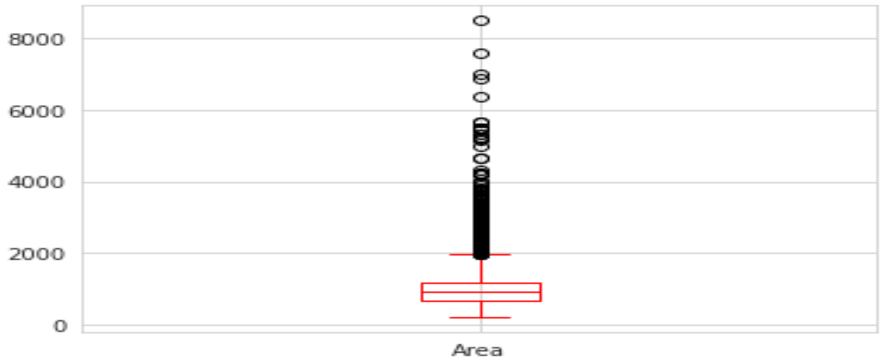


Figure 5: Uni-variate outlier

As we can see from the boxplot that was just presented to us, the typical distribution of the data is contained within the block, whilst the outliers are indicated by the small circles that are located at the very end of the graph. In this study, the researchers identified outliers by using the IQR

approach to determine limitations on the sample values that are a factor k of the IQR. Now that the outliers have been dealt with, we can plot the boxplot once more and see whether or not they have been handled. After treating the data with IQR, it has been found that there are no outliers.

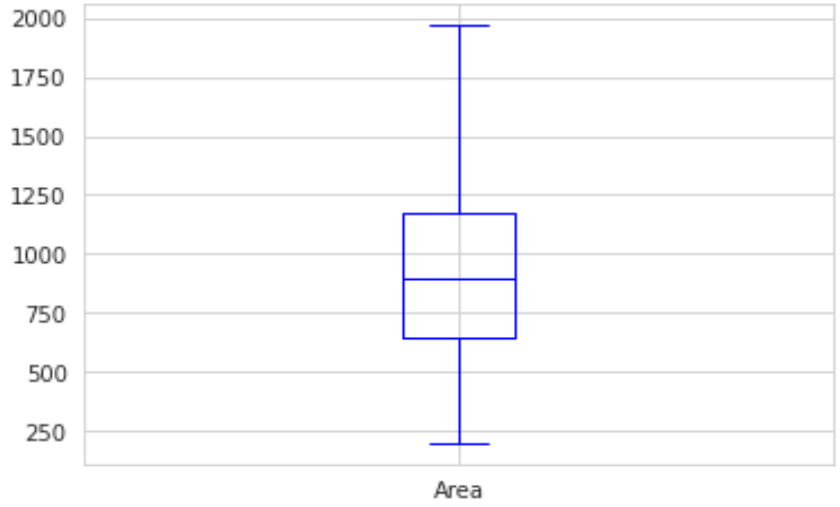


Figure 5: Handled Uni-variate outlier with IQR

Feature Importance

The relative significance of each feature contained within the training data is referred to as the feature's importance, which is an essential notion in the field of machine learning. In a nutshell, it tells us which traits are the best predictors of the variable that we are interested in. The stage in the machine learning model construction pipeline known as "determining feature importance" is one of the most important steps. Calculating the relevance of a feature may be done in a number of different ways; however, the majority of the time, these methods focus on determining some kind of score that evaluates how frequently a feature is utilised in the model as well as how much it contributes to the predictions as a whole. The relevance of features can also assist us in locating possible issues with either our data or the methodology that we use for modelling.

Sklearn's *RandomForestClassifier* is used for weighing the significance of individual features. This will be helpful in feature selection by locating the most significant features for addressing classification problems using machine learning.

Permutation Based Feature Importance

It is possible to employ the permutation-based importance to get around the disadvantages of the default feature importance, which is estimated based on the mean impurity decrease. It is referred to as the permutation importance approach in the scikit-learn software package. This method will jumbled each feature in a random order and then compute how the model's performance has

changed as a result. The most significant elements are the ones that have the most influence on the overall performance. In order to plot the significance, the eli5 library was utilized.

An ordered collection of features, along with the significance values for those characteristics, is generated by using feature permutation importance explanations. The model predictions are more sensitive to aspects of the data that are found higher up in the ranks. Features with a lower rank have a less impact on the predictions made by the model. In addition, the importance values provide a representation of the relative significance of the traits. The following is an illustration of this: The price of a house is most heavily influenced by factors such as the lot size, longitude and latitude of the property, and other similar factors.

Weight	Feature
1.1866 ± 0.2145	Area
0.2451 ± 0.1317	Longitude
0.0496 ± 0.0427	Latitude
0.0086 ± 0.0063	JoggingTrack
0.0078 ± 0.0002	SportsFacility
0.0073 ± 0.0067	No. of Bedrooms
0.0064 ± 0.0095	Intercom
0.0045 ± 0.0068	VaastuCompliant
0.0014 ± 0.0010	LiftAvailable
0.0014 ± 0.0011	ClubHouse
0.0007 ± 0.0006	PowerBackup

Figure Permutation Based Feature Importance

The first number in each row shows how much model performance decreased with a random shuffling (in this case, using "accuracy" as the performance metric).

Feature Importance With Shap

It is possible to apply the SHAP interpretation (it is model-agnostic), which may then be used to compute the feature importances derived from the Random Forest. It estimates how much each individual attribute contributes to the prediction by making use of the Shapley values that are derived from game theory. The computation of feature importances using SHAP can involve significant amounts of processing effort. Nevertheless, it is able to supply additional information such as decision plots and dependence plots. Researchers utilized the summary plot function in order to plot the relevance of features using a horizontal bar plot.

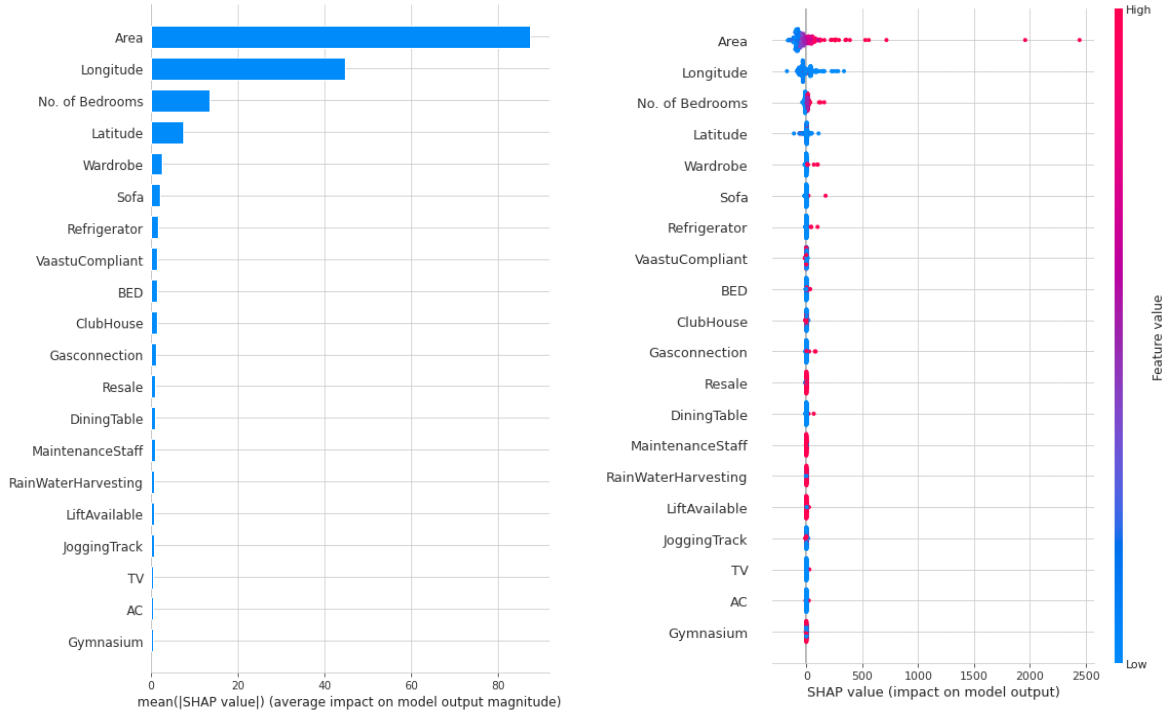


Figure Feature Importance with Shap

Longitude is the most important feature. Understandably, the area of the house, Number of bedrooms available in the house, latitude plays a major role in the final price too of the house.

Future Direction

After the EDA process is complete, the author will proceed to apply machine learning algorithms to the dataset that was created. Because of this, after the EDA process, it will be possible to get a model with improved accuracy. As a result, in the subsequent research that we do, we will design a machine learning model with the potential to improve the accuracy with which housing prices may be predicted and considerably contribute to the accurate assessment of real estate price trends. Also, the following author will execute the model in the context of the application of predicting house prices, in addition to developing a more robust algorithm that is based on a more advanced machine learning approach.

Conclusion

Previous research that was relevant to the topic of predicting house prices relied primarily on hedonic-based methods. These are typical statistical approaches that have some limitations due to the assumptions and calculations they make. In more recent studies, researchers have attempted to compare and contrast traditional methods with machine learning techniques such as neural

networks and support vector machines. On the other hand, this research makes use of a variety of EDA techniques in order to carry out an exhaustive exploratory analysis of a dataset including predictions of house prices. The process consists of a number of processes, including comprehensive data normalization, handling outliers, determining the importance of features, dealing with missing information, and so on. by examining a number of various criteria, such as the size of the house in terms of both area and square footage, the number of bedrooms it has, and so on. The researcher has utilized the folium library in this investigation in order to create an interactive base map of Mumbai, complete with the amount of homes currently available for purchase at each particular area.

References

1. Aytekin. (2010). The use of fuzzy logic in predicting house selling price. Expert Systems with Applications , 1808–1813.
2. Azadeh, A. Z. (2012). A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations. . Expert Systems with Applications , 298–315.
3. Fan, G. O. (2006). Determinants of house price. A decision treea pproach. Urban Studies , 2301–2315.
4. Gerek, L. H. (2014). House selling price assessment using two different adaptive neuro-fuzzy techniques. . Automation in Construction , 33–39.
5. Gu, J. Z. (2011). Housing price forecasting based on genetic algorithm and support vector machine. . Expert Systems with Application , 3383–3386.
6. Kauko, T. H. (2002). Capturing housing market segmentation: An alternative approach based on neural network modeling. Housing Studies , 875–894.
7. Liu, J. Z. (2006). Application of fuzzy neural network for real estate prediction . LNCS , 1187–1191.
8. Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. . Expert Systems with Applications , 2843–2852.
9. Wang, X. W. (2014). Real estate price forecasting based on SVM optimized by PSO. Optik-International Journal for Light and Electron Optics , 1439–1443.