House Rental Price Prediction System

Dhruvil Modi[1]
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
dhruvim1@umbc.edu

Abstract

In this study, we suggest developing multiple prediction models based on machine learning to determine the current data of the real estate in order to more precisely estimate the housing price or its changing trend in the future. This will allow us to analyze the impact of various factors on the housing price. The cost of the house and the ideal moment to purchase it are both known to the buyer thanks to the price. The customer is more likely to choose a home and participate in the bidding process when the pricing is suitable. We have used python libraries and techniques for the software implementation of the project. On the training data, we ran five distinct machine learning algorithms: K-Nearest Neighbor, Random forest, XG Bootsting, SVM and Linear Regression.

Keywords- Machine Learning, house price, K-Nearest Neighbour, Random Forest, AdaBoost

1. Intoduction

Real estate today serves as both a man's essential need and a symbol of their prosperity and status. Real estate investments appear to be profitable in general since their property prices don't drop off quickly. This sector attracts consumers and businesses due to the numerous profit potential it offers because to the increased need for homes globally. Numerous factors, such as demography, the economics, and politics, have an impact on these demands. This makes it difficult for data analysts and ML engineers to analyze such marketplaces around the world since they have to consider a variety of scientific disciplines, each of which deals with a distinct type of data, in order to provide consumers and stakeholders with correct results.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Since the earlier house price predictions were difficult, the best strategy is needed to obtain an accurate projection. Missing features are a challenging issue to handle in machine learning models, let alone house prediction models, and data quality is a vital factor in predicting the price of homes. As a result, feature engineering becomes a crucial technique for making models that are more accurate.

Property values typically rise over time, so it is necessary to determine their current value. This valued value is necessary for the property sale, loan application, and marketability of the property. These valued values are established by qualified evaluators. The drawback of this method is that these appraisers could be partial as a result of the interests that buyers, sellers, or mortgages have conferred. Therefore, we require an automated prediction model that can assist in accurately forecasting property values.

The rest of this paper is structured as follows. Problem statement is provided in Section 2 followed by the dataset description. Proposed approach is mentioned in Section 3. Section 4 describes the result analysis, whereas Section 5 and 6 discusses the conclusion and future work respectively.

2. Problem Definition

The project's goal is to forecast house rental prices in order to assist those who intend to purchase a home by letting them know the price range in the future so they may properly arrange their finances. The research uses machine learning to make predictions. Using several machine learning algorithms to anticipate prices reduces the amount of manual work required while improving accuracy, making the process simpler.

3. Proposed Approach

3.1. Dataset

The dataset used in this project is data on house rental prices with various characteristics in India. The dataset has a CSV (Comma-Separated Values) format and has 4746 samples with 12 features.The dataset has 4 int64 type features and 8

object-type features.

This dataset contains details on about 4700+ houses, apartments, and flats that are available to rent, along with other factors such as BHK, Rent, Size, Number of Floors, Area Type, Area Locality, City, Furnishing Status, Type of Preferred Tenant, Number of Bathrooms, and Point of Contact.

3.1.1. Exploratory Data Analysis

Initially, we loaded the dataset that is mentioned above. To understand data better, we pre-processed the data by dropping the columns that are unnecessary and removing the null values. I, then performed Univariate and Multivariate Analysis.

In Univariate analysis, I described individual columns to understand the dataset well with respect to the number of availabilities of area, tenant type, furnishing status and much more. I have then plotted the histograms for the better reference of the requisite features of the data-set.

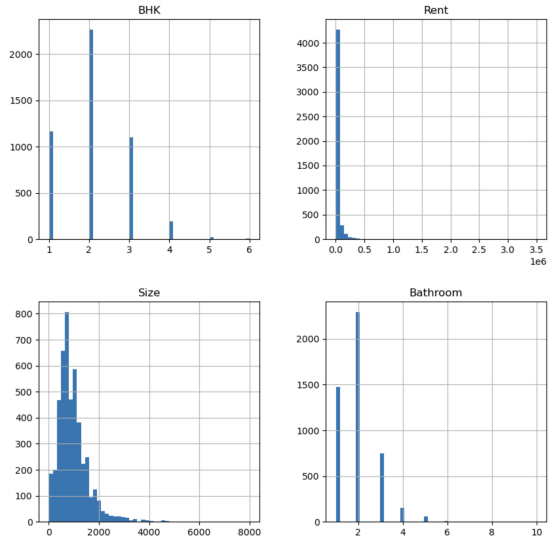

Figure 1. Histogram of requisite features

In the multivariate analysis, I have taken in consideration multiple columns at a time to get the desired column results. For instance, I generated a new column named 'Price per sqft' using the columns 'Size' and 'Rent'. Moreover, I have assumed the threshold of 300 sqft/bhk to detect the size per BHK outlier. Then I have created a function that Removes price per sqft outlier with mean and one standard deviation. Then I have checked bathroom outliers and removed unnecessary outliers.

Figure 2 displays the correlation matrix between numeric features and target features which is price in this case. Moreover, figure 3 shows the correlation plot between categorical features and target features.

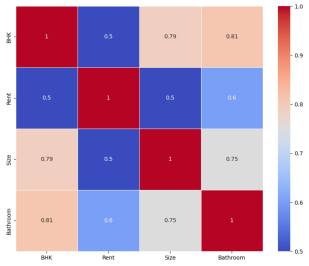

Figure 2. Correlation matrix between the features.

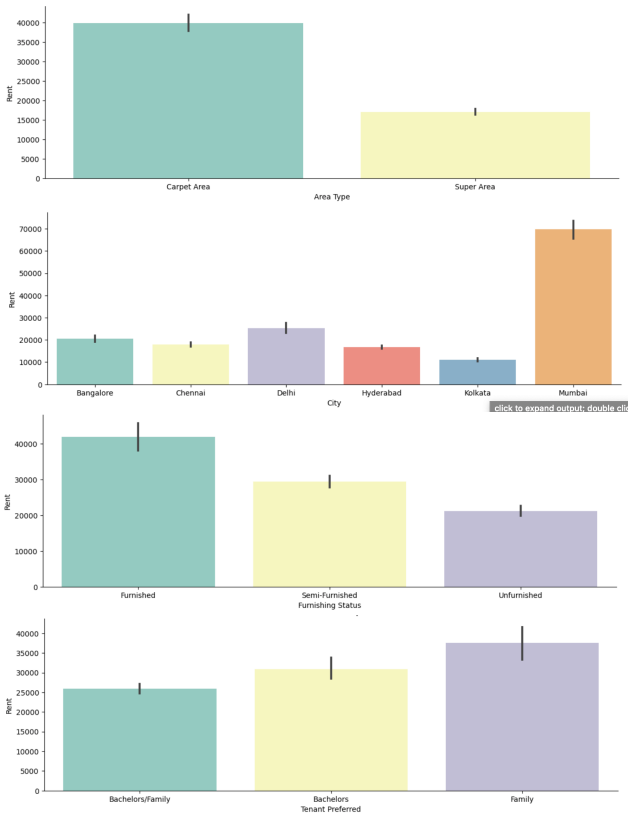

Figure 3. Correlation plot between the features.

3.2. Model used

After pre-processing and understanding the dataset, I created train and test split for the data to perform Normalization.

The five models that I have used to perform the house rental prediction are K-Nearest Neighbours, Random forest, Adaboost, Linear regression and Support Vector Machine.

3.2.1. K-Nearest Neighbour

The algorithm is supervised learning classifier that exploits the proximity to generate predictions about clustering of single data point. The parameter k in the kNN represents the number of closest neighbour that is involved in the voting process. As a part of the project implementation, I have given consecutive values between 5 and 15 to find the best score. It was noticed that the model outgave the best score when the value of k was 7. Hence, we used 7 as a parameter to find the accuracy of kNN model which was observed to be 0.727.

3.2.2. Random Forest

The algorithm is an incredibly used supervised machine learning approach to solve classification and regression issues. The algorithm uses random feature subsets and performs effectively with high dimensional dataset. While performing the model the I received the best score with the maximum depth of 8, number of estimator 25 and random state 11. Moreover, the accuracy using the parameters was found out to be 0.932.

3.2.3. Adaboost

The algorithm runs by fitting the classifier to the initial data set and then fits the additional copies of the classifier to the same data set, but with the weights of instances that were incorrectly classified being changed so that the subsequent classifier would concentrate more on complicated cases. In the implementation performed in the project I received best score when the learning rate was 0.1, Number of estimators were 100 and random state was 11. The accuracy using these parameters was observed to be 0.899.

3.2.4. Multiple Linear Regression

The algorithm analyses the relation between dependent and independent variables. In order to anticipate the value of any dependent variable, multiple linear analysis uses known values of independent variables. When applied the algorithm for our house rental prediction data set, we get the best score of 0.701. On training on X and Y variables we get the model accuracy as 0.641.

3.2.5. Support Vector Machine

SVM, also known as support vector machine, is a supervised machine learning approach that is used for classification and regression. The goal of the method is to find a hyperplane in an N dimensional space that accurately classifies the data points. The parameters that i considered for the implementation included kernal which could be either linear or sigmoid. It was found that the best score was derived when the kernal was linear. Since the SVM algorithm is used more for classification, using it for prediction system was bound to give us a low accuracy. Hence the model accuracy that we received was 0.534

	model	best_score	best_params
0	knn	0.460230	{'n_neighbors': 7}
1	boosting	0.856539	{'learning_rate': 0.1, 'n_estimators': 100, 'r...
2	random_forest	0.893655	{'max_depth': 8, 'n_estimators': 25, 'random_s...
3	linear_regression	0.701729	{}
4	SVM	0.115647	{'kernel': 'linear'}

Figure 4. Model and their Best Score

3.3. Evaluation

In this part, I implemented Mean Squared Error to evaluate our model. Figure 5 shows the Mean Squared Error on train and test data for each model.

	train	test
KNN	157941.525431	655749.486951
RF	63498.211157	163275.610997
Boosting	171831.281366	243784.09386
LinearRegression	351929.983399	862722.853276
svm	1256652.46645	2688190.525953

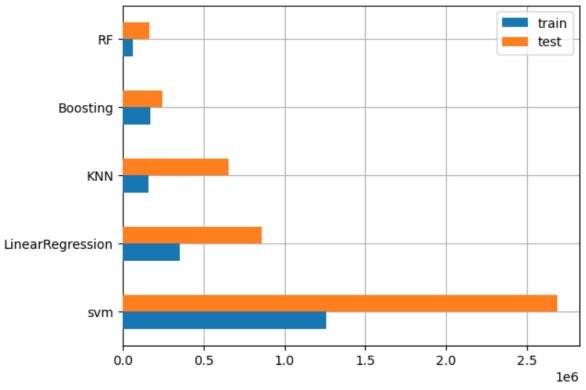

Figure 5. Model and their Best Score.

4. Result

As a result, I have successfully derived the prediction for house rental for each model. According to figure 6, we can observe the housing price prediction by different model.

	y_true	prediction_KNN	prediction_RF	prediction_Boosting	prediction_LinearRegression	prediction_svm
1733	11000	11071.4	10913.1	13168.5	2672.0	15766.4
1442	13000	12071.4	13297.7	13168.5	14944.0	16029.9
1911	18000	21571.4	22240.4	13517.4	24592.0	16082.4
2003	22000	23428.6	17893.1	18242.7	25232.0	16235.2
553	11000	16500.0	11114.7	13168.5	6576.0	15894.3

Figure 6. Model and their Best Score.

5. Conclusion and Future work

5.1. Conclusion

This study examines various housing price prediction methods. Some distinct categories of machine learning techniques are K-Nearest Neighbor, Random forest, XG Bootsting, SVM and Linear Regression. Despite the fact that all those techniques produced pleasing outcomes, various models each have advantages and disadvantages. From the research survey and project implementation, it can be concluded that Random Forest classifier proves to be the most accurate and efficient model in prediction house pricing depending on the given features. Although the Random Forest approach is prone to overfitting, it has the lowest error on the training set. Due to the necessity of fitting the dataset several times, it has a significant time complexity.

5.2. Future Work

As a future scope of the project, it is recommended to learn more about these models, especially the combinations of several models, more research on the following subjects shall be done:

- The impact of numerous regression models coupled together.
- The capacity of machine learning models to ”re-learn.”
- The fusion of deep learning and machine learning techniques.
- The underlying reasons for tree-based models' successful performance.
- The easier methods for fitting complex models.

References

Arietta, Sean M., e. a. City forensics: Using visual elements to predict non-visual city attributes. In *IEEE transactions on visualization and computer graphics 20.12 (2014)*, 2014.

Banerjee, D. and Dutta., S. Predicting the housing price direction using machine learning techniques. In *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI). IEEE, 2017*, 2017.

Fan C, Cui Z, Z. X. Housing price prediction with ,achine learning algorithms. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing - ICMLC 2018.*, 2018.

Lakshmi, B. N. and Raghunandhan, G. H. A conceptual overview of data mining. In *National Conference on Innovations in Emerging Technology. IEEE, 2011*, 2011.

MoreiradeAguiar, M., S. R. a. B. A. Housingmarketanalysis using a hierarchical–spatial approach: the case of belo horizonte, minas gerais, brazil. In *Regional Studies, Regional Science, 1(1)*.

Sabyasachi Basu, T. G. T. Analysis of spatial autocorrelation in house prices. In *The Journal of Real Estate Finance and Economics 17.1*, 1998.

Sifei Lu, Zengxiang Li, Z. Q. X. Y. R. S. M. G. A hybrid regression technique for house prices prediction. In *IEEE International Conference on Industrial Engineering and Engineering Management(IEEM), Singapore*, 2017.

Steven C. Bourassa, Eva Cantoni, M. E. R. H. S. D. Housing submarkets and house price prediction. In *The Journal of Real Estate Finance and Economics*, 2007.

Visit Limsombunchai, Christopher Gan, M. L. House price pre- diction: hedonic price model vs. artificial neural network. In *New Zealand Agricultural and Resource Economics Society Conference.*, 2004.