Rental Price Prediction

Akshi Bharadwaj
Central University of Rajasthan
2021MSBDA004

Niharika Jha
Central University of Rajasthan
2021MSBDA025

Nikesh Kumar Tiwari
Central University of Rajasthan
2021MSBDA026

***Abstract*—Today, it's hard to find a property that isn't out of your price range. On top of that, the current interest rates make it a challenge to purchase a house right now. It is safe to say that the real estate market is going through a tough time. In this paper, we have built a few models to predict the rent prices of houses and compared the accuracy of each to understand which works better. Decision Tree, K-nearest neighbors and Random Forest are trained on a dataset from Mumbai and compared by calculating Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-squared values. We have also used some interesting concepts such as feature engineering, one hot encoding, and standardization before applying these models.**

Keywords: Rent prices, machine learning, feature engineering, one hot encoding, kNN, decision tree, random forest.

I. Introduction

Rental price prediction is a process of forecasting the rental price of a property based on its characteristics and the local rental market. [1] Due to individual differences, everyone has diverse requirements, budgets and options for renting a house. [7]The prediction of rental prices can help landlords, property managers, and tenants make informed decisions about renting properties. For landlords and property managers, accurate rental price predictions can help them set competitive prices for their properties and attract potential tenants. For tenants, rental price predictions can provide valuable information about the rental market, which can help them make informed decisions about where to rent and how much to pay. This paper is an attempt to demonstrate various machine learning algorithms for price prediction. In recent years, machine learning has introduced alternative econometric approaches for fitting and predicting house/rental prices [9]. Chen et al. (2016) summarized two main advantages of machine learning methods over traditional statistical methods [10]. First, statistical methods usually make strong assumptions concerning the randomness of measure errors of data (e.g., normal distribution), which may not always follow the real-world situation. Second, machine learning methods are able to capture high-order interactions among features and non-linear relationships between dependent and independent variables. Recent developments in deep learning additionally provide ways to better utilize unstructured data such as texts, images, and audio [11].

II. Literature Survey

A. Paper name: Real Estate Price Prediction with Regression and Classification (2016) [2]

The dataset used is the prices and features of residential houses sold from 2006 to 2010 in Ames, Iowa, obtained from the Ames Assessor's Office. This dataset consists of 79 house features and 1460 houses with sold prices. This paper used two types of supervised learning algorithms- classification and regression. In order to determine the regularization parameter, throughout the project in both classification and regression parts, k-fold cross-validation with k = 5 on a wide range of selection of regularization parameters was performed. Further, PCA was performed on all models to improve results. It was concluded that the area per square foot, the material of the roof, and the neighbourhood have the greatest statistical significance in predicting a house's sale price.

B. Paper name: Real Estate Price Prediction Using Machine Learning [3]

This paper used the real estate housing data taken from the UCI machine learning repository. The data is spread across 20000 rows and has ten attributes. In this paper, models like the random forest, SVM, multiple regression, gradient boosting and multi-layer perceptron are used. The ensemble learning algorithm is used to reduce variance and improve accuracy. Then the concept of mean absolute error is used to measure accuracy. To sum it up, the random forest has more accuracy in prediction when compared to other methods used in the study like multiple regression, Support vector machine, gradient-boosted trees, neural networks, and bagging.

C. Paper Name: House Price Forecasting using Machine Learning [4]

The datasets for real estate properties are collected from various real estate websites which included attributes like location, carpet area, built-up area, age of the property, zip code, etc. This paper presents the results of the prediction of the market value of a real estate property which aims at applying the ideas of machine learning on how to enhance and foresee the costs with high accuracy in the fluctuating pricing of the real estate market. The paper proposes a system that predicts house prices using a regression machine learning algorithm. In the paper, the Decision tree machine learning algorithm is used to construct a prediction model to predict

potential selling prices for any real estate property. The system provides 89% accuracy while predicting the prices for real estate prices.

D. Paper Name – Real Estate Price Prediction using correlation analysis and multiple linear Regression Analysis. [5]

The dataset used in this model has been taken from the sold house in Zhaoqing City in southern China from the time period 2010 -2020. The objective of this project was to get an accurate prediction of house pricing in Zhaoqing city using correlation analysis and multiple linear regression Analysis. The house price prediction in Zhaoqing city included some external factors such as GDP, tertiary industry, Government policies, the income of urban residents etc. First, the methods of correlation analysis were used, and variables that are highly correlated with house price data were selected based on correlation coefficients. Then, the model was constructed for predicting the house price on the basis of multiple linear regression analyses conducted with selected variables. The Pearson coefficient correlation analysis and multiple linear regression analysis were chosen for this research. Finally, the value of the goodness-of-fit and the difference between the predicted and actual house prices of house prices were combined to observe the prediction effect. Hence, from the above dataset, using correlation Analysis and multiple linear regression analysis, the price of real estate in Zhaoqing city of China was predicted.

III. Methodology

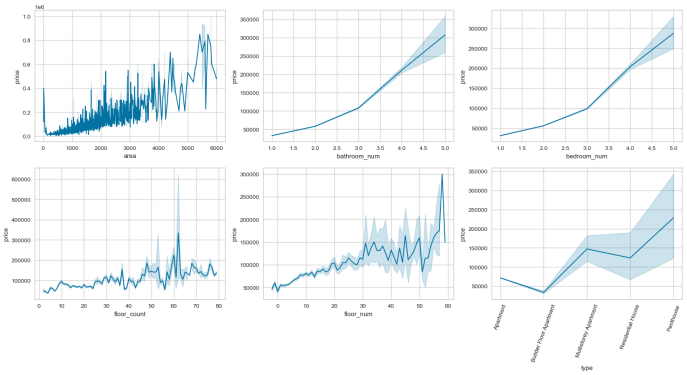
A. Data collection and EDA

The dataset was imported from the online data science community Kaggle [6] in Comma Separated Value (CSV) format.

The columns in the original data were:

'area', 'bathroom_num', 'bedroom_num', 'city', 'desc', 'dev_name', 'floor_count', 'floor_num', 'furnishing', 'id', 'id_string', 'latitude','locality', 'longitude', 'post_date', 'poster_name', 'price', 'project', 'title', 'trans', 'type', 'url', 'user_type'

Exploratory Data Analysis (EDA) is an approach to analyzing and understanding a dataset. It is an important step in the machine learning workflow, as it can help identify patterns, relationships, and trends in the data. Python libraries like Matplotlib, Seaborn and Yellowbricks were used for visualization. The following graph shows a general trend between the target and feature variables.



B. Preprocessing

Data preprocessing is the process of preparing the data for analysis and modeling. It is an important step in the machine learning workflow, as it can greatly affect the performance of the models. The Python programming language was used and Pandas and NumPy were used for data processing. Since the data comprised of the information of just one city, i.e., Mumbai, latitude and longitude do not play an important role, so they were removed. Similarly, any missing or irrelevant information was removed to clean the data and make it suitable for use in a machine learning model. Additionally, our data comprised categorical variables; we applied one hot encoding to convert it into numerical variables. Then the data was normalized using Sklearn's Normalizer.

C. Training and Testing

Two divisions were made of the entire dataset- 80% of the data was for training and the remaining 20% for testing. We used Sklearn's train_test_split to split the data. K-nearest neighbors, Decision Tree and Random Forest were applied to train the data. The trained data is then applied to test dataset to predict the prices.

D. Metrics

We applied several Machine Learning techniques to train the model and calculated their corresponding Root Mean Squared Error (RMSE), Mean Absolute Error(MAE) and R-squared score to evaluate the model. The scikit-learn library was used to implement the Machine Learning algorithms.

Root Mean Squared Error (RMSE) is a way to measure the error in the model, typically used in regression analysis. Lower values of RMSE indicate a better fit of the model to the data.

The formula for Root Mean Squared Error (RMSE) is:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2}$$

where,
n is the number of observations,
p_i is the predicted value for observation i

o_i is the actual value for observation i

Root Mean Squared Percentage Error (RMSPE) is a measure of the difference between predicted values and actual values, expressed as a percentage.The formula for calculating RMSPE is:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(p_i - o_i)^2}{p_i}}X100$$

where,
n is the number of observations,
p_i is the predicted value for observation i
o_i is the actual value for observation i

Mean Absolute Error(MAE): The Mean Absolute Error (MAE) measures the average of the absolute differences between the predicted values and the true values. It tells us how close the predicted values are to the true values on average. It tells us how close the predicted values are to the true values and lower values of MAE indicate that the model's predictions are closer to the true values on average. The formula for Mean Absolute Error (MAE) is:

$$\frac{1}{n}\sum_{i=1}^{n}|p_i - o_i|$$

where,
n is the number of observations,
p_i is the predicted value for observation i
o_i is the actual value for observation i

Mean Absolute Percentage Error(MAPE) expresses the error as a percentage of the actual values. It is calculated as the average absolute difference between the predicted values and the actual values, divided by the actual values and multiplied by 100. MAPE is a variation of MAE, where it is expressed as a percentage of the actual values. The formula for Mean Absolute Error (MAE) is:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|p_i - o_i|}{p_i}X100$$

where,
n is the number of observations,
p_i is the predicted value for observation i
o_i is the actual value for observation i

R-sqaured: R-squared, also known as the coefficient of determination, is a statistical measure that tells us how well a model fits the data. It ranges between 0 and 1, where 0 means the model explains none of the variability of the response data around its mean, and 1 means the model explains all the variability of the response data around its mean. It compares the predicted values from a model to the true values and calculates the proportion of variance in the dependent variable that can be explained by the independent variable.

It's important to note that a high R-squared value does not necessarily mean that the model is a good fit for the data, as it does not take into account other factors such as overfitting or outliers. Additionally, R-squared is not a measure of model predictions accuracy, instead, it's a measure of how well the model fits the data, and it's a measure of how well the model describes the relationship between the predictor and response variables. The formula for R-squared is:

$$R^2 = \frac{SS_{residual}}{SS_{total}}$$

where:
$SS_{residual}$ is the sum of squared residuals (i.e., the sum of the squared differences between the predicted values and the true values), and
SS_{total} is the total sum of squares (i.e., the sum of the squared differences between the true values and the mean of the true values)

E. Models

a) ***k-nearest Neighbours (kNN)****:* kNN is a simple and effective machine learning algorithm that can be used for both classification and regression problems. It is a non-parametric algorithm, meaning it does not assume any functional form. kNN is also considered a lazy algorithm since the algorithm doesn't actively learn from the training data. At training time, all it is doing is storing the complete data set but it does not do any calculations at this point. All the computation happens when we apply the model on unseen data points.

An appropriate value of k plays an important role in the accuracy of the model. We did hyperparameter optimization using CV Grid Search to find an optimal value of k. However, it can be computationally expensive, especially when the training set is large.

After the value of k is decided, distances between the point we want to predict and all the training points are calculated. There are multiple methods to calculate distance. Two of the most popular methods are as follows:

Euclidean distance: square root of the sum of the squared differences between a new point (x) and an existing point (y). The formula is as follows:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan distance: sum of the absolute differences between a new point (x) and an existing point (y). The formula is as follows:

$$\sum_{i=1}^{k}|x_i - y_i)|$$

For regression, the distances are then sorted in ascending order and average of 'k' minimum distances (or, median) is then considered the predicted value.

kNN is also sensitive to the scale of the predictors, so it is important to standardize the predictors before using kNN regression.

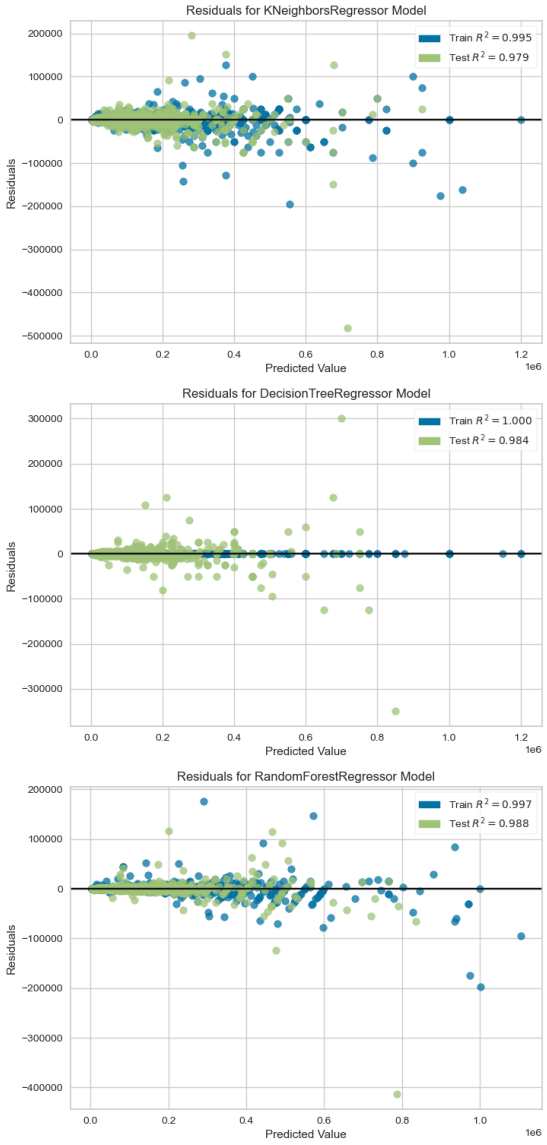
*b) **Decision Tree**:* [13] Decision Tree (DT) is used to solve regression and classification problem. It splits the data into small subsets and evaluated. Both continuous and categorical outcome features can be worked by decision tree. It is a flowchart-like structure in which the internal node represents a feature (or attribute), the branch represents a decision and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It has the ability to observe the non-linear correlation of independent attributes with dependent target attributes.
Decision tree regressor considers some features as input and trains the model to estimate the future outcome feature.
One of the methods for splitting the node used when the target variable is continuous is **reduction in variance**. It is so-called because it uses variance as a measure for deciding the feature on which node is split into child nodes.
For each split, the variance of each child node is individually calculated and the split with the lowest variance is selected.
*c) **Random Forest**:* Random Forest is an ensemble learning method for both classification and regression problems. In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. [12] Random Forest has the same concept like Decision Tree by reproducing a large number of trees in forest then this algorithm will restart training sample and randomly choose features and observe to build decision tree and choose in improving the accuracy. Hence, the random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn randomly from a training set with replacement, called the bootstrap sample. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. The final prediction is made by averaging the predictions of the individual trees. It is also less sensitive to the scale and distribution of the predictors. Random forest can handle high-dimensional data with correlated predictors, it also deals well with missing data. [8] Howard and Bowles (2012) claim "ensembles of decision trees (often known as 'Random Forests') have been the most successful general-purpose algorithm in modern times." However, it can be computationally expensive, especially when the number of trees is large, and it may be difficult to interpret the results of a random forest model.

IV. Conclusion

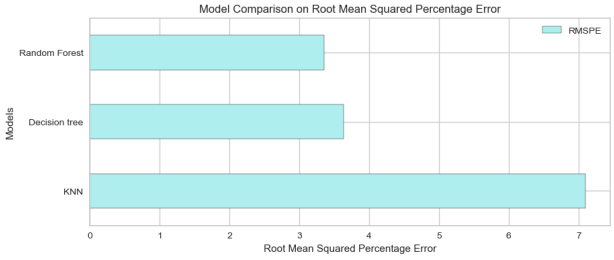
The following table shows the RMSE Scores and accuracy scores of the dataset for different machine learning algorithms:

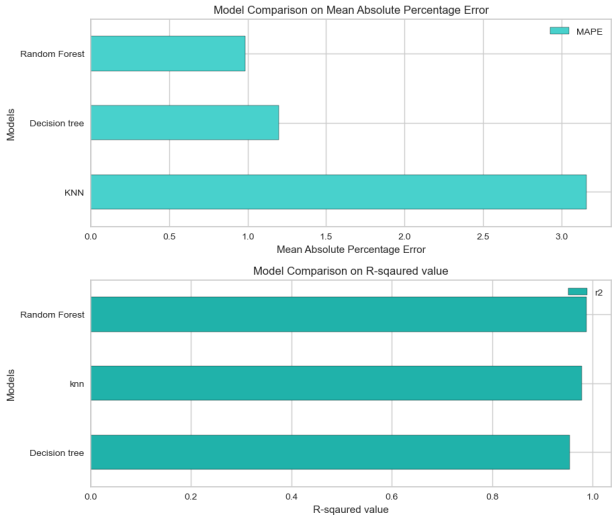
Algorithms	RMSPE	MAPE	R^2	Accuracy
Random forest	3.36	0.98	0.98	0.98
K-nearest neighbors	7.09	3.16	0.97	0.97
Decision tree	4.09	1.23	0.97	0.98

Following are the residual plots for each model:



After training and testing of datasets with all models, the highest R-squared value and least error value are achieved by the Random Forest Classifier closely followed by K-nearest neighbors.
Following is a comparison of models w.r.t. RMSPE, MAPE and R^2 values, respectively:





This paper examined and analyzed different machine learning algorithms used to predict rent prices. An accurate prediction model would allow tenants to know what prices to expect as well as the homeowners to put a realistic price tag on their property. Everything considered the results of the project have shown the potential of kNN, Decision Tree and Random Forest in predicting house prices.

REFERENCES

[1] Preston, V., Taylor, S.M. Personal Construct Theory and Residential Choice, Annals of the Association of American Geographers, 1981.

[2] Yu, H., & Wu, J. (2016). Real estate price prediction with regression and classification. CS229 (Machine Learning) Final Project Reports.

[3] Ravikumar, Aswin Sivam. "REAL ESTATE PRICE PREDICTION USING MACHINE LEARNING". Diss. Dublin, National College of Ireland, 2017.

[4] Kuvalekar, A., Manchewar, S., Mahadik, S., & Jawale, S. (2020, April). House Price Forecasting Using Machine Learning. In Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST).

[5] Chen, N. (2022). House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis. Wireless Communications and Mobile Computing, 2022.

[6] https://www.kaggle.com/datasets/jedipro/flats-for-rent-in-mumbai

[7] Kumar, A. (2019). House Rent Price Prediction.

[8] Howard, J., Bowles, M. (2012, February). The two most important algorithms in predictive modeling today. In Strata Conference presentation, February (Vol. 28).

[9] Mullainathan, S., Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.

[10] Chen, Y., Liu, X., Li, X., Liu, Y., Xu, X. (2016). Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning. Applied Geography, 75, 200-212.

[11] Yang, T., Xie, J., Li, G., Mou, N., Li, Z., Tian, C., Zhao, J. (2019). Social media big data mining and spatio-temporal analysis on public emotions for disaster mitigation. ISPRS International Journal of Geo-Information, 8(1), 29.

[12] Ja'afar, N. S., Mohamad, J. (2021). Application of Machine Learning in Analysing Historical and Non-Historical Characteristics of Heritage Pre-War Shophouses. Journal of the Malaysian Institute of Planners, 19(2), 72-84.

[13] W. Loh, "Classification and regression trees," WIREs Data Min. Knowl. Discov., vol. 1, no. February, pp. 1-14, 2011.