UITS

Future will be better than thy past

An initiative of PHP Family

University of Information Technology & Sciences

PROJECT PAPER

Machine Learning Lab

CSE-436

Submitted By:

Fahim Kamal

1914551046

Sec: 7A2, Batch: 45

Date: 17:07:2022

Submitted To:

Mr. Al-Imtiaz

Assistant Professor

Department of Computer Science and Engineering, UITS

Study on Flat Rent Prediction Using Machine learning Regression Algorithms

Fahim Kamal
Department of Computer Science and Engineering
University of Information Technology and Sciences
fahimkamal49@gmail.com

Mr. Al-Imtiaz
Assistant Professor
Department of Computer Science and Engineering
University of Information Technology and Sciences

Abstract:

Determining the rent price of the flat is very important nowadays as the price of the flat increases every year. So our future generation needs a simple technique to predict the flat rent in future. The rent of flat helps the people who looks for flat to know the rent of the flat. The right flat rent of the house helps the people to elect the flat to rent. There are several factors that affect the rent of the flat such as the Flat Size, Area, Total Bedrooms Number etc. This paper uses various regression techniques to predict the house price such as Linear Regression, Lasso, Decision Tree regression techniques etc.

Keywords: Flat rent prediction, Linear Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Logistic Regression, Gaussian Naïve Bayes.

Introduction:

In this study, several methods of prediction were compared to finding the best predicted results in determining a flat's rent compared to the actual rent. This paper brings the latest research on regression technique that can be used flat rent prediction such as Linear regression, Lasso regression, Tree Regression, Random Forest Regression, Logistic Regression and Gaussian Naïve Bayes. As the initial flat rent prediction were challenging and require some best method to get accurate prediction.

Data quality is a key factor to predict the flat rents and missing features are a difficult aspect to handle in machine learning models let alone flat rent prediction model. In general, the value of the property increases over time and its valued value must be calculated. During the sale of property or while applying for the loan and the marketability of the property, this valued value is required. The professional evaluators determine these valued values. However, the disadvantage of this practice is that these evaluators could be biased because buyers, sellers or mortgages have bestowed interest.

Apartment rental prices are influenced by various factors. The aim of this study is to analyze the different features of an apartment and predict the rental price of it based on multiple factors. I have made the dataset from taking data from bProperty.com which includes the rental price and different features of apartments in the city of Dhaka, Bangladesh. The results show the accuracy and prediction of the rent of an apartment, also indicates the different types of categorical values that affect the machine learning models. Another purpose of the study is to find out the factors that signify the apartment rental price in Dhaka. [1]

A dataset of approximately 1462 houses entries including 14 features has been collected from bProperty.com for our research. The dataset is to be divided between an appropriate ratio for training and testing purpose.

Related Works:

There have been various extensive researches conducted on Flat rent prediction through the use of Machine Learning. Most of these works have been done in the context of developed countries. Although the house rental prices in Bangladesh are not very systematic, a pattern is present and we hope to evaluate the factors that affect the prices. There are some work done in the field of flat rent pricing of Bangladesh. Previous works use various methods to implement a model for this type of study [2]. One of the most popular ways to predict house pricing through machine learning is the use of Linear Regression as the model contains many features affecting the price [3]. Then there are many algorithms like Lasso Regression, Decision Tree Regression, Random Forest Regression XGBoost (XGBT) which has been used to predict house prices in California [4]. An approach to the use of Artificial Neural Network was used to predict the house prices in New Zealand [5]. It proved to be a daunting task as the multiple feature required powerful calculations from algorithms, but the results were promising.

Dataset:

This dataset of mine consists of consists of 1461 rows and 15 columns or attributes. They are: Area, Flat Size, Nearby Bazar, Floor Number, Total Bedrooms, Attached Washrooms, Total Balconies, Floor Type, Garage Facilities, Lift Service, Gas Availability, Security Service, CCTV Coverage, Walking Distance From Main Road, Rent.

A. **Area:** It means the location where the flat resides in. The flat rent varies depending on the location of the flat.

B. **Flat Size:** It means the total size of the flat sq ft. It’s one of the most crucial attribute in the dataset.

C. **Nearby Bazar:** It means the location of the flat is near bazar or it has some distance from the location.

D. **Floor Number:** It defines the number of floor of the flat.

E. **Total Bedrooms:** It defines the number of bedrooms in the flat.

F. **Attached Washrooms:** It defines the number of attached washrooms in the flat.

G. **Total Balconies:** It defines the number of total balconies in the flat.

H. **Floor Type:** It means the design of the floor in the flat.

I. **Garage Facilities:** It means if the flat has garage facilities or not.

J. **Lift Service:** It means if the flat has lift service available or not.

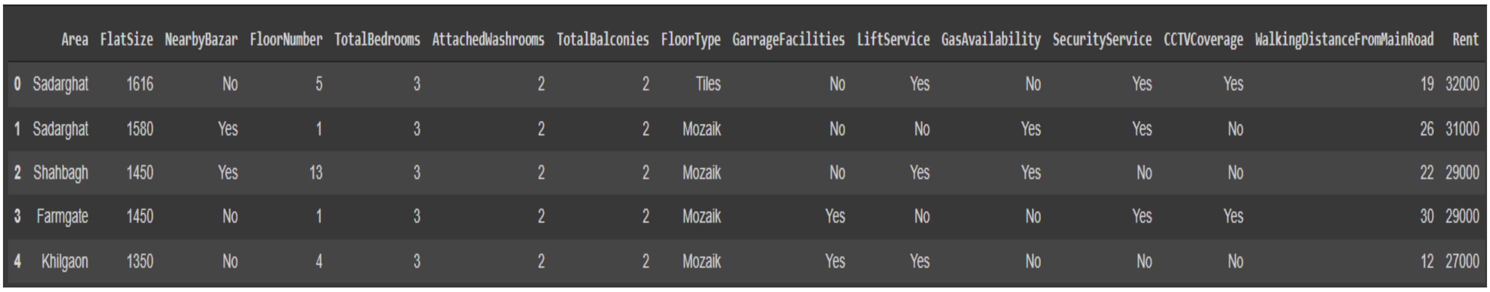
K. **Gas Availability:** It means if the flat has gas available or not.

L. **Security Service:** It means if the flat has security service available or not.

M. **CCTV Coverage:** It means if the flat has CCTV coverage available or not.

N. **Walking Distance From Main Road:** It defines the walking distance from main road from the flat.

O. **Rent:** It defines the rent price. It is the dependent value of the dataset.



	Area	FlatSize	NearbyBazar	FloorNumber	TotalBedrooms	AttachedWashrooms	TotalBalconies	FloorType	GarrageFacilities	LiftService	GasAvailability	SecurityService	CCTVCoverage	WalkingDistanceFromMainRoad	Rent
0	Sadarghat	1616	No	5	3	2	2	Tiles	No	Yes	No	Yes	Yes	19	32000
1	Sadarghat	1580	Yes	1	3	2	2	Mozaik	No	No	Yes	Yes	No	26	31000
2	Shahbagh	1450	Yes	13	3	2	2	Mozaik	No	Yes	Yes	No	No	22	29000
3	Farmgate	1450	No	1	3	2	2	Mozaik	Yes	No	No	Yes	Yes	30	29000
4	Khilgaon	1350	No	4	3	2	2	Mozaik	Yes	Yes	No	No	No	12	27000

Project dataset’s top five rows

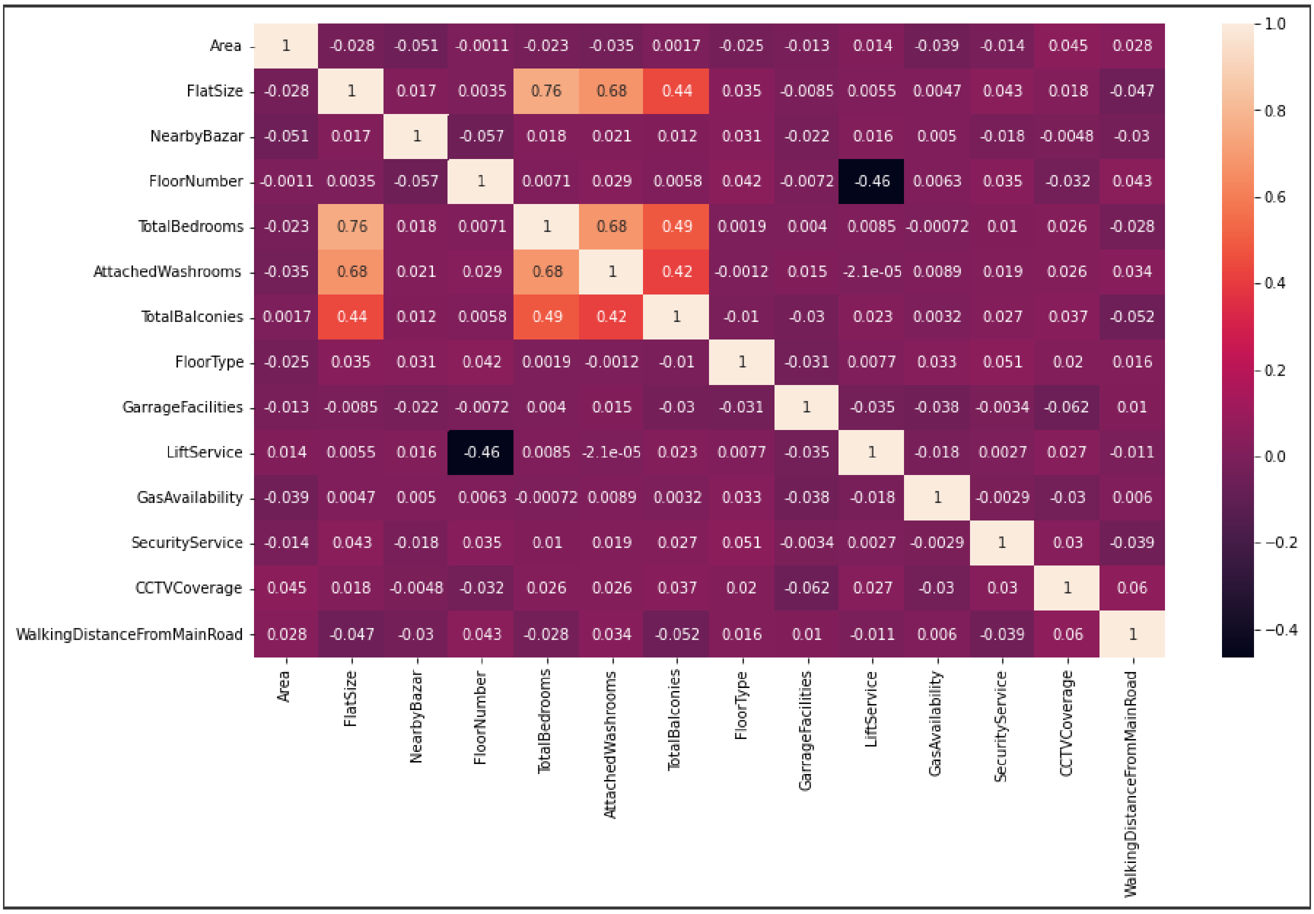


Fig. Feature correlations for the dataset

Data Source:
https://www.bproperty.com/

Dataset Github Link:
https://github.com/DarkraiFahim/FlatRentPrediction.git

2. Regression Problems:

Regression Problems also known as Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc.

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling and determining the causal-effect relationship between variables.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, "Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum." The distance between datapoints and line tells whether a model has captured a strong relationship or not. [6]

There are many regression models like Linear Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression Logistic Regression, Gaussian Naive Bayes in our data set.

2.1. Linear Regression:

It is a commonly used algorithm and can be imported from the Linear Regression class. For finding a relationship between two continuous variables, Linear regression is useful. One variable is predictor or independent, and the other variable is variable response or dependent. It looks for a relationship that is statistical but not deterministic. It is said that the relationship between two variables is deterministic if the other can express one variable accurately.

$y=b*x + c$

where Y is the dependent variable, X is the independent variable. Theta is the coefficient factor. When you use more than one independent variable to get output, it is termed Multiple linear regression. This kind of model assumes that there is a linear relationship between the given feature and output, which is its limitation. [7]

2.2. Lasso Regression:

Lasso (Least Absolute Shrinkage and Selection Operator), similar to Ridge Regression, also penalizes the absolute size of the coefficients of regression. It is also capable of reducing variability and enhancing linear regression models accuracy. The regression of Lasso differs from the regression of the ridge in a way that uses absolute values instead of squares in the penalty function. This leads to a penalization (or equivalent limitation of the sum of the absolute values of the estimates) which causes some estimates of the parameters to turn out to be exactly zero. [7]

2.3. Decision Tree Regression:

It is a non-linear Machine Learning algorithm. It breaks down a data set into smaller and smaller subsets by splitting resulting in a tree with decision nodes and leaf nodes. Here the idea is to plot a value for any new data point connecting the problem. The kind of way in which the split is conducted is determined by the parameters and algorithm, and the split is stopped when the minimal number of information to be added reaches. Decision trees often yield good results, but even if any slight change in data occurs, the whole structure changes, meaning that the models become unstable. [8]

2.4. Random Forest Regression:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. The diagram above shows the structure of a Random Forest. You can notice that the trees run in parallel with no interaction amongst them. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model. [9]

2.5. Logistic Regression:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives.

A logistic regression model can take into consideration multiple input criteria. In the case of college acceptance, the logistic function could consider factors such as the student's grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into one of two outcome categories. [10]

2.6. Gaussian Naïve Bayes:

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. We have explored the idea behind Gaussian Naive Bayes along with an example.

When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution.

An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label, which is all what is needed to define such a distribution. [11]

Table: without feature selection

Regression	R^2 Score	Mean squared error	Explain variance score	Mean absolute percentage error
Linear Regression	0.9295	2577281.553	0.9295	0.0149
Lasso Regression	0.9296	2575245.92	0.9296	0.0148
Decision Tree Regression	0.9136	3159873.084	0.9141	0.0108
Random Forest Regression	0.8639	4976356.653	0.8656	0.0083
Logistic Regression	0.8739	4611774.744	0.874	0.0433
Gaussian Naïve Bayes	0.9584	1519624.573	0.9587	0.0044

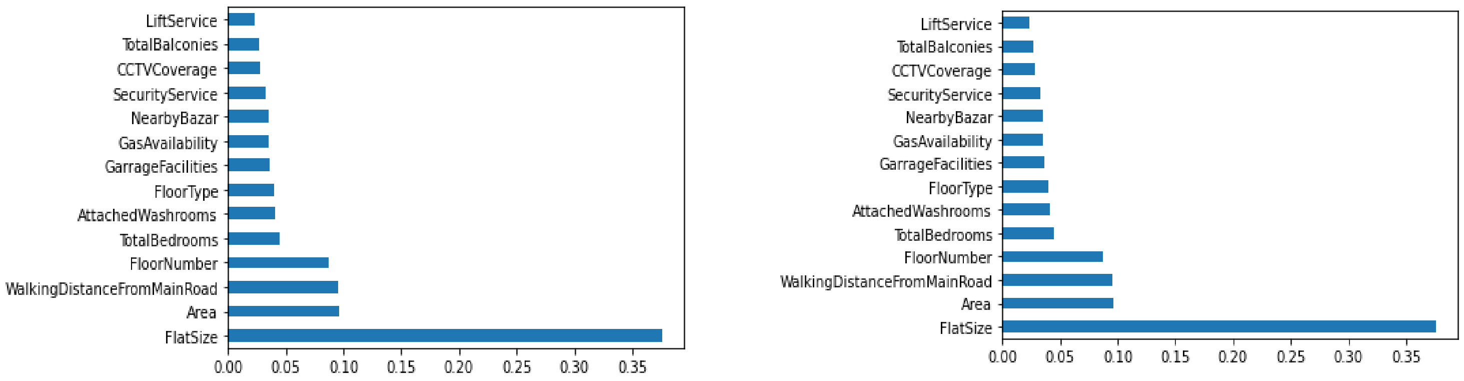


Fig: Feature Selection for Random Forest, Decision Tree

Top 3 Regression with feature selection

Regression	Number of feature	R² Score	Mean squared error	Explain variance score	Mean absolute percentage error
Gaussian Naïve Bayes	11	0.9614	1410409.5563	0.9615	0.0042
Gaussian Naïve Bayes	8	0.9613	1413822.5256	0.9614	0.0044
Gaussian Naïve Bayes	5	0.9283	2622013.6519	0.9287	0.0057
Lasso Regression	11	0.9299	2563986.8458	0.9299	0.0148
Lasso Regression	8	0.9299	2563839.8419	0.9299	0.0147
Lasso Regression	5	0.929	2594705.9101	0.9291	0.0142
Linear Regression	11	0.9298	2565360.5524	0.9299	0.0149
Linear Regression	8	0.9299	2565104.8053	0.9299	0.0148
Linear Regression	5	0.929	2596452.1468	0.929	0.0143

Validation:

Regression	Cross Validation Score	Initial Score
Linear Regression	0.8973	0.9295
Lasso Regression	0.8973	0.9296
Decision Tree Regression	0.8444	0.9141
Random Forest Regression	0.9101	0.8656
Logistic Regression	0.1442	0.874
Gaussian Naïve Bayes	0.9388	0.9587

Linear Regression: In the Linear Regression we are using python k-fold cross validation. In the dataset we have total 15 attributes, after the feature selection we are just using 7 feature and we got the minimum accuracy as 81% and maximum accuracy as 95% using linear model. The mean score is 90%.

Lasso Regression: In the Lasso Regression we are using python k-fold cross validation. In the dataset we have total 15 attributes, after the feature selection we are just using 7 feature and we got the minimum accuracy as 81% and maximum accuracy as 95% using linear model. The mean score is 90%.

Decision Tree Regression: In the Decision Tree Regression we are using python k-fold cross validation. In the dataset we have total 15 attributes, after the feature selection we are just using 7 feature and we got the minimum accuracy as 61% and maximum accuracy as 96% using linear model. The mean score is 84%.

Random Forest Regression: In the Random Forest Regression we are using python k-fold cross validation. In the dataset we have total 15 attributes, after the feature selection we are just using 7 feature and we got the minimum accuracy as 76% and maximum accuracy as 98% using linear model. The mean score is 91%.

Logistic Regression: In the Logistic Regression we are using python k-fold cross validation. In the dataset we have total 15 attributes, after the feature selection we are just using 7 feature and we got the minimum accuracy as 13% and maximum accuracy as 16% using linear model. The mean score is 14%.

Gaussian Naïve Bayes: In the Gaussian Naïve Bayes we are using python k-fold cross validation. In the dataset we have total 15 attributes, after the feature selection we are just using 7 feature and we got the minimum accuracy as 91% and maximum accuracy as 96% using linear model. The mean score is 94%.

Result:

R² Score:

The R2 score is one of the performance evaluation measures for regression-based machine learning models. It is also known as the coefficient of determination. The R2 score is a very important metric that is used to evaluate the performance of a regression-based machine learning model. It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset. Simply put, it is the difference between the samples in the dataset and the predictions made by the model. [12]

$$R^2 = 1 - SS_{res} / SS_{tot}$$

Where,

SS_{res} is the sum of squares of the residual errors.

SS_{tot} is the total sum of the errors.

The R2 score ranges from 1, a perfect score, to negative values for under-performing models. The scores that you can achieve and their meaning can be seen here:

i. A score of 1 is the perfect score and indicates that all the variance is explained by the independent variables.

ii. A score of 0 would indicate that the independent variables don't explain any of the variance.

iii. A negative score below 0 indicates that the independent variables aren't explaining the variance and are actually contributing negatively to the model.

An important reminder when looking at the R2 scores from different models is that the variance found in a dataset is not comparable across datasets, meaning that R2 scores can not be used to directly compare model performance. [13]

We can see, R^2 value differs for every algorithm used in this model. But the best score came out for Gaussian Naïve Bayes. It's score is 0.9584. The second one is Lasso Regression which score is 0.9296 and the third one is Linear Regression and it's score is 0.9295.

Mean Squared Error:

In statistics, the mean squared error (MSE) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors. [14] Here, the error is the difference between the attribute which is to be estimated and the estimator. The mean square error may be called a risk function which agrees to the expected value of the loss of squared error. This difference or the loss could be developed due to the randomness or due to the estimator is not representing the information which could provide a more accurate estimate.

The mean squared error can also be referred to the second moment of the error, measured about the origin. It includes both the variance and bias of the estimator. If an estimator is an unbiased estimator, then its MSE is the same as the variance of the estimator. The unit of MSE is the same as the unit of measurement for the quantity which is being estimated. [15]

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

There is no correct value for MSE. Simply put, the lower the value the better and 0 means the model is perfect. Since there is no correct answer, the MSE's basic value is in selecting one prediction model over another. [16]

We can see, Mean squared error value differs for every algorithm used in this model. But the less score came out for Gaussian Naïve Bayes. It's mean squared error is 1519624.573. The second one is Lasso Regression which error is 2575245.92 and the third one is Linear Regression and it's error is 2577281.553.

Explain Variance Score:

In machine learning, variance is the difference between the actual samples of the dataset and the predictions made by the model. When working on a regression-based machine learning problem, it is very useful to know how much of the variance is explained by the machine learning model.

The explained variance is used to measure the proportion of the variability of the predictions of a machine learning model. Simply put, it is the difference between the expected value and the predicted value. It is a very important concept to understand how much information we can lose by reconciling the dataset. It is used on regression-based problems to measure the difference between samples and predictions.

$$explained\ variance(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

In simple words, Explain Variance Score is the proportion of the variability of the difference between the actual samples of the dataset and the predictions made by the model. While using it, always remember that the concept of Explained Variance is used to measure the proportion of the variability of the predictions of a regression based machine learning model. [17]

Mean Absolute Percentage Error:

Mean Absolute Percentage Error is an error metric used to measure the performance of regression machine learning models. It is a popular metric to use amongst data scientists as it returns the error as a percentage, making it both easy for end users to understand and simpler to compare model accuracy across use cases and datasets.

MAPE returns error as a percentage, making it refreshingly easy to understand the 'goodness' of the error value. The lower the percentage, the more accurate the model, so 10% is better than 60%. It goes without saying that how 'good' your MAPE score is depends very much so on your use case and dataset, but a general rule of thumb that I follow is:

MAPE	Interpretation
<10 %	Very good
10 % - 20 %	Good
20 % - 50 %	OK
> 50 %	Not good

The mathematical formula for MAPE is as follows:

$$\mathrm{M} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

MAPE is a popular metric to use for regression machine learning models, however there are some things one must consider when optimizing for this metric. It gives error as percentage and it is very easy for user to understand. [18]

Conclusion:

The model structure developed in this paper produces significant benefits for rent prediction in Dhaka, enabling an inference of conceptual rent at some major places in a Dhaka city. This results in for us to get an approximate idea about flat rent in Dhaka city.

Future applications of the model structure should provide significant benefits to the valuation process, and better understanding the potential impact on rents, or value, of proposed new construction. As a result, it should also provide benefits to the property management process.

Reference:

[1] Asif Ahmed Neloy, H M Sadman Haque and Md. Mahmud Ul Islam, "Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring", *ResearchGate*, Feb 2022. Available: https://www.researchgate.net/publication/333150583_Ensemble_Learning_Based_Rental_Apartment_Price_Prediction_Model_by_Categorical_Features_Factoring

[2] Temilola Aderibigbe and Hongmei Chi, "Investigation of Florida Housing Prices using Predictive Time Series Model", *ACM Digital Library*, July 22, 2022. Available: https://dl.acm.org/doi/10.1145/3219104.3229253

[3] R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher, "Real estate value prediction using multivariate regression models", *IOPScience*. Available: https://iopscience.iop.org/article/10.1088/1757-899X/263/4/042098

[4] Byeonghwa Park and Jae Kwon Bae, "Using machine learning algorithms for housing price prediction", *ScienceDirect*, April 15, 2022. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417414007325

[5] New Zealand Agricultural and Resource Economics Society(2004), "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network", *EconPapers*. Available: https://econpapers.repec.org/paper/agsnzar04/97781.htm

[6] Javatpoint, "Regression Analysis in Machine learning", *Javatpoint*. Available: https://www.javatpoint.com/regression-analysis-in-machine-learning

[7] Adarsh Kumar, "House Rent Price Prediction*", International Research Journal of Engineering and Technology*, April, 2019. Available: https://www.irjet.net/archives/V6/i4/IRJET-V6I4677.pdf

[8] Surabhi S, "A Quick Overview of Regression Algorithms in Machine Learning", *Analytics Vidhya*, January 10, 2021. Available: https://www.analyticsvidhya.com/blog/2021/01/a-quick-overview-of-regression-algorithms-in-machine-learning/

[9] Chaya Bakshi, "Random Forest Regression", *Level Up Coding*, June 9, 2020. Available: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

[10] George Lawton, Ed Burns and Linda Rosencrance, "Logistic Regression", *TechTarge*t. Available: https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

[11] Prateek Majumder, "Gaussian Naive Bayes", *OpenGenus*. Available: https://iq.opengenus.org/gaussian-naive-bayes/

[12] Aman Kharwal, "R2 Score in Machine Learning", *THECLEVERPROGRAMMER*, June 22, 2021. Available: https://thecleverprogrammer.com/2021/06/22/r2-score-in-machine-learning/#:~:text=The%20R2%20score%20is%20one%20of%20the%20performance,squared%20score%2C%20then%20this%20article%20is%20for%20you.

[13] Stephen Allwright, "What is a good R2 (R-Squared) score and how do I interpret it?", *Stephen Allwright*, April 16, 2022. Available: https://stephenallwright.com/good-r2-score/

[14] Moshe Binieli, “Machine learning: an introduction to mean squared error and regression lines”, *freeCodeCamp*, October 16, 2018. Available: https://www.freecodecamp.org/news/machine-learning-mean-squared-error-regression-line-c7dde9a26b93/

[15] BYJUS, “Mean Square Error-Definition and Formula”, *BYJUS*. Available: https://byjus.com/maths/mean-squared-error/

[16] Walker Rowe, “Mean Square Error & R2 Score Clearly Explained”, *bmc blogs*, July 5, 2018. Available: https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/

[17] Aman Kharwal, “Explained Variance in Machine Learning”, *THECLEVERPROGRAMMER*, June 25, 2021. Available: https://thecleverprogrammer.com/2021/06/25/explained-variance-in-machine-learning/

[18] Stephen Allwright, “What is a good MAPE score?”, *Stephen Allwright*, October 27, 2021. Available: https://stephenallwright.com/good-mape-score/