Predict the Rent of House in Dhaka City by Applying Machine Learning Techniques

Submitted By:

Name: Md. Saif Ahammod Khan
Id: 1721779042
Email: saif.ahammod@northsouth.edu
Department of Electrical and Computer Engineering,
North South University, Dhaka 1229, Bangladesh

Name: Rasheeq Ishmam
Id: 1831350042
Email: rasheeq.ishmam@northsouth.edu
Department of Electrical and Computer Engineering,
North South University, Dhaka 1229, Bangladesh

Name: S. M. Sajid Hasan Shanta
Id: 1831238642
Email: sajid.shanta@northsouth.edu
Department of Electrical and Computer Engineering,
North South University, Dhaka 1229, Bangladesh

Submitted To:
Dr. Sifat Momen
Associate Professor
Department of Electrical and Computer Engineering,
North South University, Dhaka 1229, Bangladesh

Course Details:

Course Title: Machine Learning
Course Code: CSE445
Section: 1
Date of Submission: 01-05-2022

Predict the Rent of House in Dhaka City by Applying Machine Learning Techniques

Abstract

Several factors influence the cost of renting a house. The goal of this research is to look at the many features of a house and anticipate the rental price based on a variety of parameters. We have used an online housing platform, BProperty, to collect 38190 pieces of house rental data from Dhaka city for visual analysis and prediction. The results demonstrate the accuracy and predictability of a house's rent, as well as the many sorts of categorical data that influence machine learning models.

I. Introduction

An accurate prediction on the house rent is important to prospective homeowners, developers, investors, appraisers, tax assessors, and other real estate market participants, such as mortgage lenders and insurance companies. Therefore, the availability of a housing rent prediction model fills an important information gap and improves the general public's scenario.

Dhaka, the capital of Bangladesh, is one of the world's most densely populated megacities. The high rate of in-migration, territorial expansion, and natural growth have all contributed to Dhaka's rapid population rise. It boosts people's desire for housing, and as a result of this need, house rents have been rising substantially day by day. For a long time, the city's general population has been severely oppressed by the property owners. House rents are set by landlords without regard for any regulations. There should be a standard criterion for determining house rent that will prevent landlords from acting aggressively.

Here we scraped a dataset of 38190 data samples from an online housing property based in Dhaka listed on BProperty. Then we collect and merge more data from the Google Map integrated with BProperty. With this dataset, we build a Machine Learning model to predict the house rent of Dhaka using Linear Regression, Ridge regression, LASSO Regression, Decision Tree, Random Forest Regressor, and Extreme Gradient Boosting.

Several factors influence the cost of renting an apartment. The goal of this research is to look at the many features of an apartment and anticipate the rental price based on a variety of parameters.

II. Literature review

Several in-depth studies on house rent prediction have been conducted with Machine Learning. The majority of these studies have been conducted in developed countries. Although housing rental prices in Bangladesh are not particularly systematic, they do follow a pattern, and we intend to identify the elements that influence them.

In [1] Neloy et al. utilized the Advanced Regression Techniques (ART) and established an acceptable model to predict the rental price by comparing different features of an apartment using BProperty.com data containing 3505 samples for training and evaluating the model. After cleaning the data, the authors used the Advance Linear Regression, Neural Network, Random Forest, Support Vector Machine (SVM), and Decision Tree Regressor algorithms as the base predictors. The Ensemble learning was stacked on the following algorithms – Ensemble AdaBoosting Regressor, Ensemble Gradient Boosting Regressor, Ensemble XGBoost. Also, Ridge Regression, Lasso Regression, and Elastic Net Regression had been used to combine the advanced regression techniques. Their highest accuracy obtained was 88.75 % using Ensemble Gradient Boosting and the lowest accuracy was 82.26% acquired from Ensemble AdaBoosting.

In [2] Yue Ming et al. used the XGBoost Algorithm to generate an effective model for predicting Chengdu Housing rental prices. For training and evaluating the algorithm, the authors scraped 36392 data samples from an online housing website. They were left with 33111 samples of original data after cleaning the dataset. They utilize the correlation coefficient matrix to determine the influence of each variable on pricing in order to analyze the elements that have a significant impact on rent. To fit the data, the RandomForestRegressor, XGBoost, and LightGBM models were used. The XGBoost algorithmic model produced the maximum accuracy of 85%, while RandomForestRegressor gave the lowest accuracy of 83%.

In [3] Embaye et al. used Ridge, LASSO, Tree, Bagging, Random Forest, and Boosting methods to the prediction of the rental value of housing over Ordinary Least Squares (OLS) methods accounting for spatial autocorrelations using household-level survey data from Uganda, Tanzania, and Malawi, across multiple years. On the other hand, Tree regression underperformed relative to the various OLS models, over the same data sets.

In [4] Yoshida et al. used regression-based and machine learning-based approaches (extreme gradient boosting (XGBoost), random forest, and deep neural network) to predict apartment rent. The results showed that, as the sample size increased, XGBoost and RF outperformed NNGP with higher out-of-sample prediction accuracy. XGBoost achieved the highest prediction accuracy for all sample sizes and error measures in both logarithmic and real scales and for all price bands.

In [5] Fei et al. scraped 8 datasets with 75250 data samples from an online housing property based in California listed on Airbnb. The authors implemented regression models with Ridge Regression, Ridge Regression with K means, SVR (support vector machine regression), Random Forest, and XGBoost. They evaluated the performance by R2, RMSE(the root of mean square error), and MAE (mean absolute error). Although random forest outperforms XGBoost in terms of R2 in the training dataset, XGBoost beats random forest in terms of R2, RMSE, and MAE while taking less time.

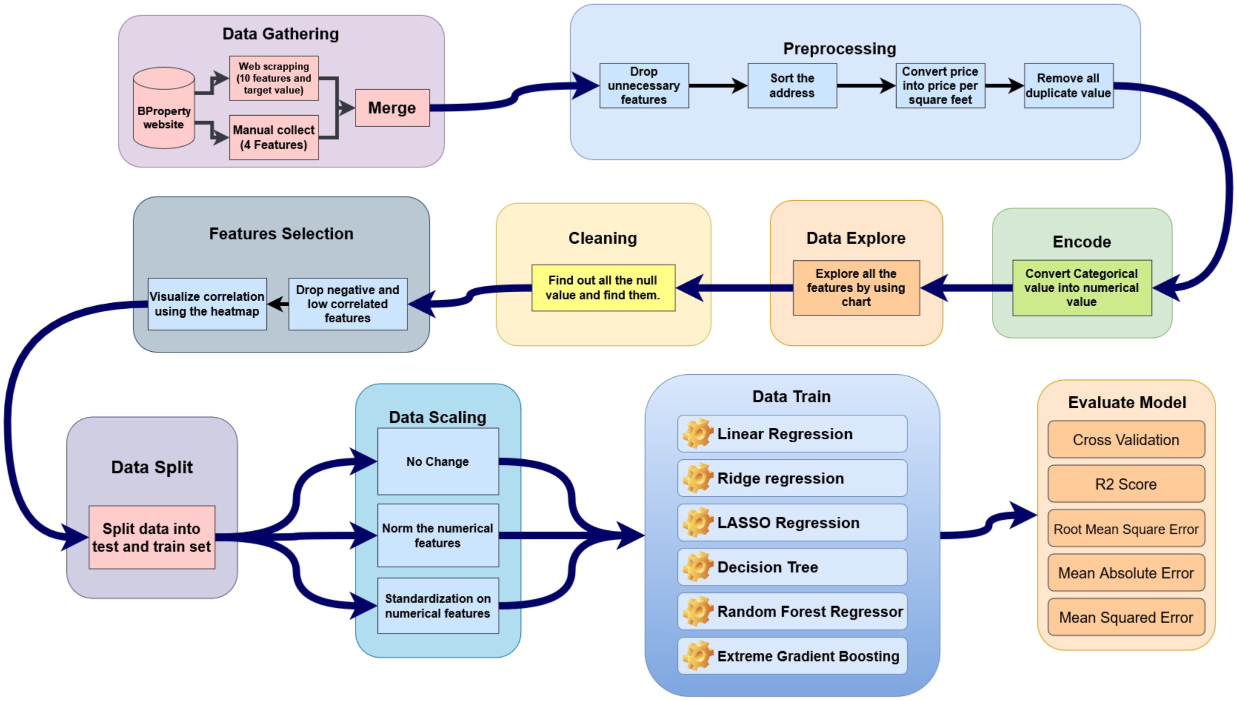
III. Methodology

The major purpose of this project is to develop a machine learning algorithm-based model. As a result, being able to forecast house rent will be beneficial. This section's data is divided into two parts, one of which was gathered using web scraping and the other which was gathered manually and afterward integrated. This work is based on a regression analysis. Different regression algorithms were used on the data in order to determine the optimum technique for predicting rent.

A. Flow Chart

This flow chart represents the workflow of the methodology.

Figure 1: Workflow of the methodology


Data Gathering
BProperty website
Web scrapping (10 features and target value)
Manual collect (4 Features)
Merge
Preprocessing
Drop unnecessary features
Sort the address
Convert price into price per square feet
Remove all duplicate value
Encode
Convert Categorical value into numerical value
Data Explore
Explore all the features by using chart
Cleaning
Find out all the null value and find them.
Features Selection
Drop negative and low correlated features
Visualize correlation using the heatmap
Data Split
Split data into test and train set
Data Scaling
No Change
Norm the numerical features
Standardization on numerical features
Data Train
Linear Regression
Ridge regression
LASSO Regression
Decision Tree
Random Forest Regressor
Extreme Gradient Boosting
Evaluate Model
Cross Validation
R2 Score
Root Mean Square Error
Mean Absolute Error
Mean Squared Error

B. Data Gathering

The data-gathering phase of this work was the most difficult. In Dhaka, there was no existing data set on dwelling rent information. As a result, data collecting for this project started from the beginning. The information was gathered from the BProperty website. BProperty is a website that allows people to purchase and rent properties all throughout Bangladesh.

Data is collected in two stages for this project. Data were acquired in the first phase utilizing a data scraping program called "Instant Data Scraper." The first dataset was prepared using the data scraping tool. There were 38189 data points in the initial dataset, 10 features, and one target value, which was the price. There are five characteristics in the second dataset: 'Address, ''Educational Institute, 'Restaurants, 'Medical Service, 'Parks.' These characteristics were carefully gathered from the BProperty website. There are 896 distinct addresses. The number of educational institutes, restaurants, medical services, and parks in the vicinity of each location were carefully tallied and then recorded in an excel file next to each address.

The only thing these two datasets have in common is the word 'Address.' Then, with the address as the common characteristic, these two datasets were merged and converted into a single dataset. This dataset contains 38189 data, 14 features, and one target value which is price. Later on, the price was changed into the price per square feet for more accurate results.

Table 1: Description of attributes

Variable Name	Description	Data Type
Brief	Brief description of the house.	string
Price	Monthly rent of the house in BDT	integer
Address	Brief address of the house	string
Type	Which type of house it is eg: Apartment	string
Details	A little bit details about the house	string
Beds	How many bedrooms are there	integer
Baths	How many bathrooms are there	integer
Size	House size in square feet	integer
Image Link	House picture link	string
Image Link 2	House picture link	string
Educational Institute	The number of educational institutes like school, college, madrasa, and university are in that area.	integer
Restaurants	The number of restaurants or food stalls are there in that area.	integer
Medical Service	The number of medical services like hospitals, pharmacies, clinics, etc. is in that area.	integer
Parks	The number of parks and playgrounds in that area.	integer

C. Data Preprocessing

Following the data collection, some per-processing work was done to eliminate any extraneous information. Because this work plan primarily uses regression methods, the optimal method is analysis. As there will be no natural language processing, no precise information is required. As a result, the 'Brief' and 'Details' functions are unnecessary. 'Image Link' and 'Image Link 2' are two additional functionalities. Deep Learning or a comparable algorithm is required to process photos. Which is not the project's aim. As a result, the variables 'Image Link' and 'Image Link 2' are no longer used.

There are 'Address' features in this variable. Although 'Address' characteristics are brief, they have been reduced to decrease complexity and just the name of the area has been maintained. It will aid in the encoding of data.

Initially, the job was done using price as the target value. As a result, all algorithms have a significant error rate. Price per square foot is afterward computed as a new goal value for greater precision. As a consequence, each data set's price per square foot is computed, and the price column is afterward removed.

There were a number of variables that appeared many times in the data. All of the duplicate values were deleted during per-processing, leaving just one unique value. The data volume was also lowered by removing the duplicate value. There were 38189 data before the duplicate value was removed, but now there are only 17076 data left, with just 10 variables remaining, one of which is the price per square foot the target value.

Table 2: Description of attributes after preprocessing

Variable Name	Description	Data Type
PricePerSqft	Monthly rent of the house in BDT	integer
Address	Brief address of the house	string
Type	Which type of house it is eg: Apartment	string
Beds	How many bedrooms are there	integer
Baths	How many bathrooms are there	integer
Size	House size in square feet	integer
Educational Institute	The number of educational institutes like school, college, madrasa, and the university is in that area.	integer
Restaurants	The number of restaurants or food stalls are there in that area.	integer
Medical Service	The number of medical services like hospitals, pharmacies, clinics, etc. is in that area.	integer
Parks	The number of parks and playgrounds in that area.	integer

D. Data Encode

When using a regression technique in machine learning, it is best to transform all category data to numerical values. When it comes to improving performance, it's a win-win situation. A numerical value is processed much faster than a textual value. This is why data encoding is required. In data encoding, each of the distinct category's variables is assigned a tag. The tag begins with a zero. A numerical value tag is assigned to each distinct categorization value.

Table 3: Categorical Value

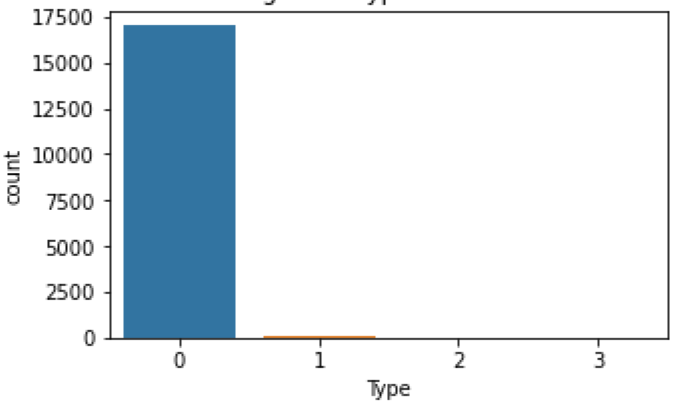
Variable Name	Data Type
Address	string
Type	string

The variables 'Address' and 'Type' in the data collection have categorical values. The name of the area is given in the address. After converting the address to a numerical value, each unique address is assigned a one-of-a-kind integer number that begins at zero. There are four varieties of houses for 'Type,' and each type is replaced with a unique numeric value. So that all of the Regression Algorithms may function properly.

E. Exploratory Work

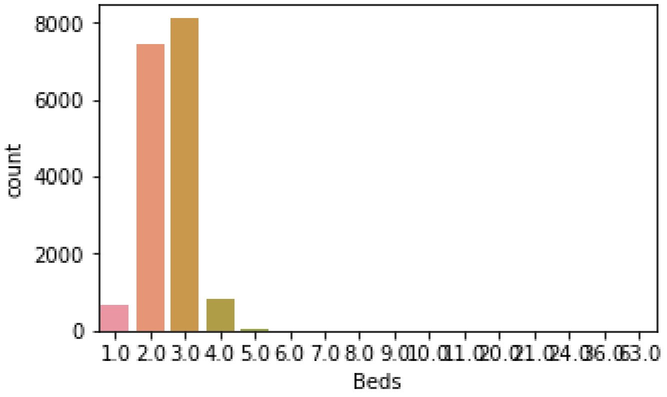
When all the features are visualized with different charts, Dhaka city's house trend can be found. Here to explore the data Count plot, Boxplot, Distribution Plot and Reg plot are used. At first individual data are explored using the count plot.

Figure 2: Type of houses



A count plot is shown in Figure 2. The various housing kinds are counted and presented here. Here, 0 represents an apartment, 1 represents a building, 2 represents a duplex, and 3 represents a penthouse. The bulk of the residences on the count site are rental apartments. There are extremely few rental buildings available, and the quantity of duplexes and penthouses is nearly non-existent.

Figure 3: Number of Bedrooms



The number of bedrooms in the rented home is depicted in Figure 3. We can observe from the graph that the majority of houses have three and two bedrooms, four bedrooms and one bedroom are few, and a five-bedroom house is nearly non-existent.

Figure 4: Number of Bathrooms

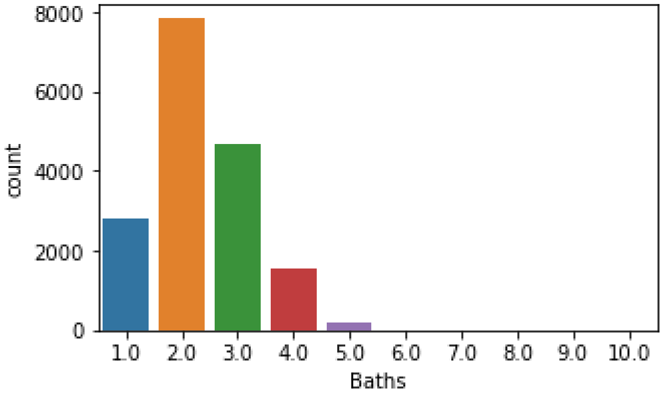


Figure 4 depicts the number of bathrooms in the rental residence. Around 8000 houses have two bathrooms, 4700 houses have three bathrooms, 3000 dwellings have one bathroom, and there are about 200 houses with five bathrooms.

Figure 5: Number of Educational Institute

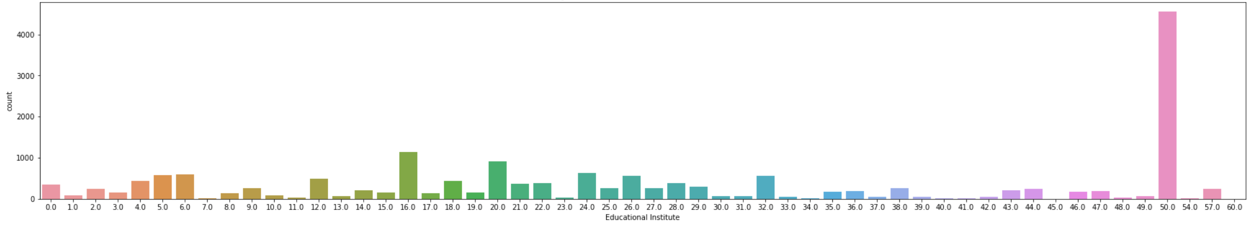


Figure 5 depicts the number of educational institutions. The university, medical college, national college, polytechnic institute, college, school, and madrasa are examples of educational institutions.

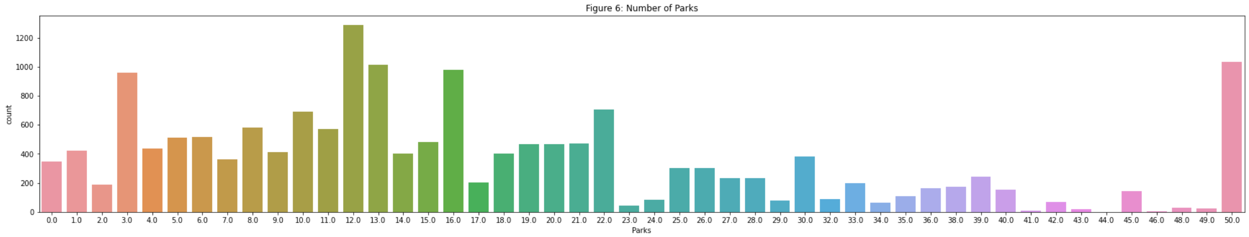


Figure 6: Number of Parks

Figure 6 shows the total number of parks. Parks include recreational areas, playgrounds, and tourist attractions.

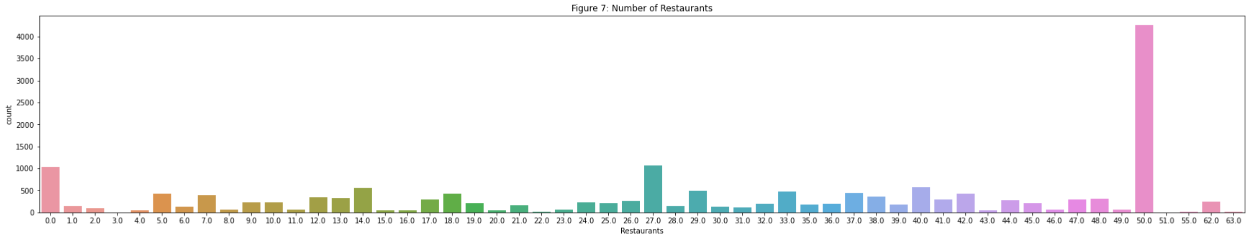


Figure 7: Number of Restaurants

The number of restaurants is shown in Figure 7. Around 1000 houses have no restaurant in their area.

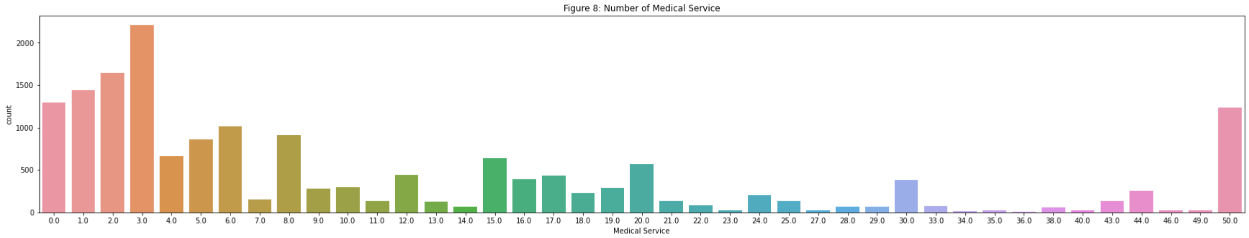
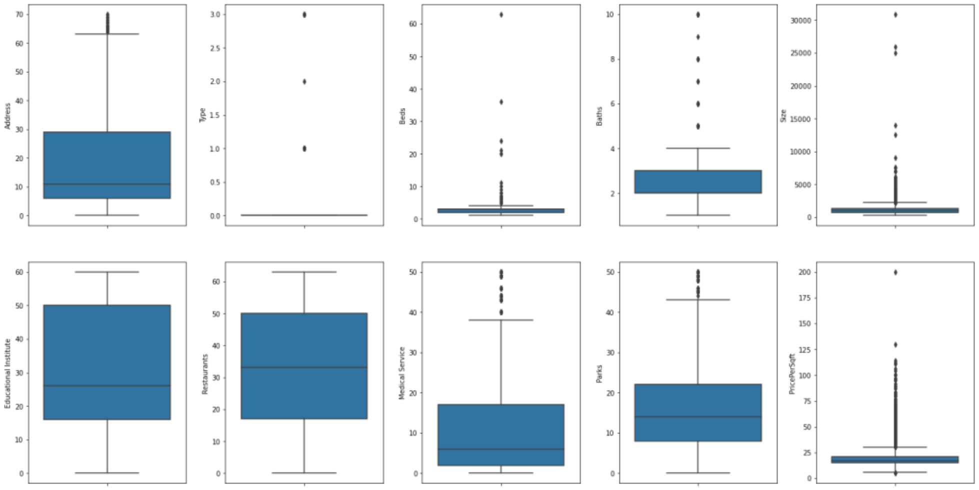


Figure 8: Number of Medical Service

The entire number of medical services is shown in Figure 8. Public hospitals, private hospitals, clinics, and medication pharmacies are all examples of medical services. Around 1300 homes do not have access to medical care, whereas 5000 homes have access to one to three medical services.

Figure 9: Boxplot to visualize all Data



Boxplot explains the minimum, maximum, median, lower quartile, and upper quartile. Here from the box plot, we can see the minimum, maximum, median, lower quartile, and upper quartile of every feature. Here every feature except Educational Institute and Medical Service has a lot of outer value.

Figure 10: Using distort to visualize all data

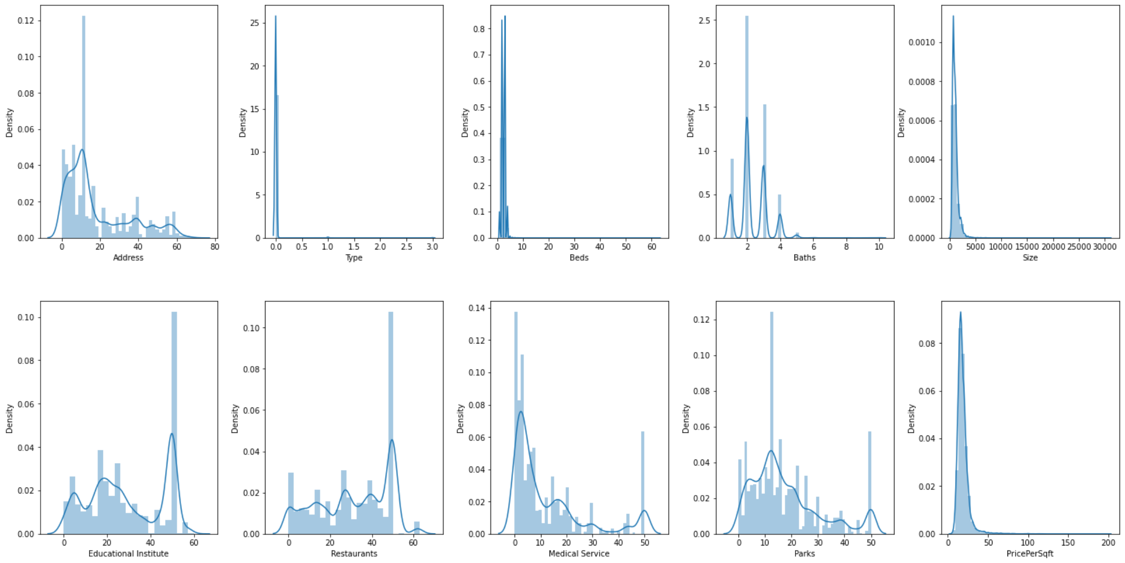
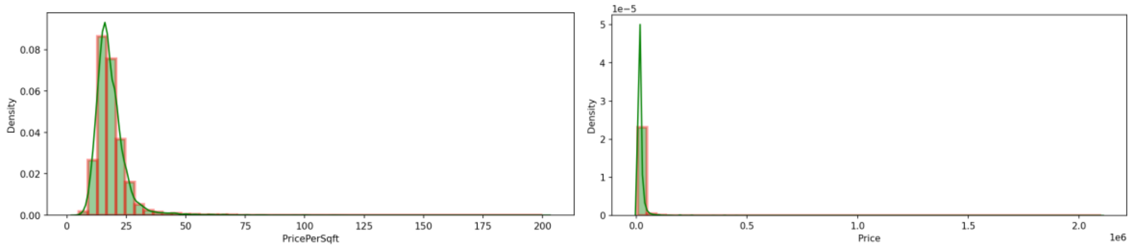


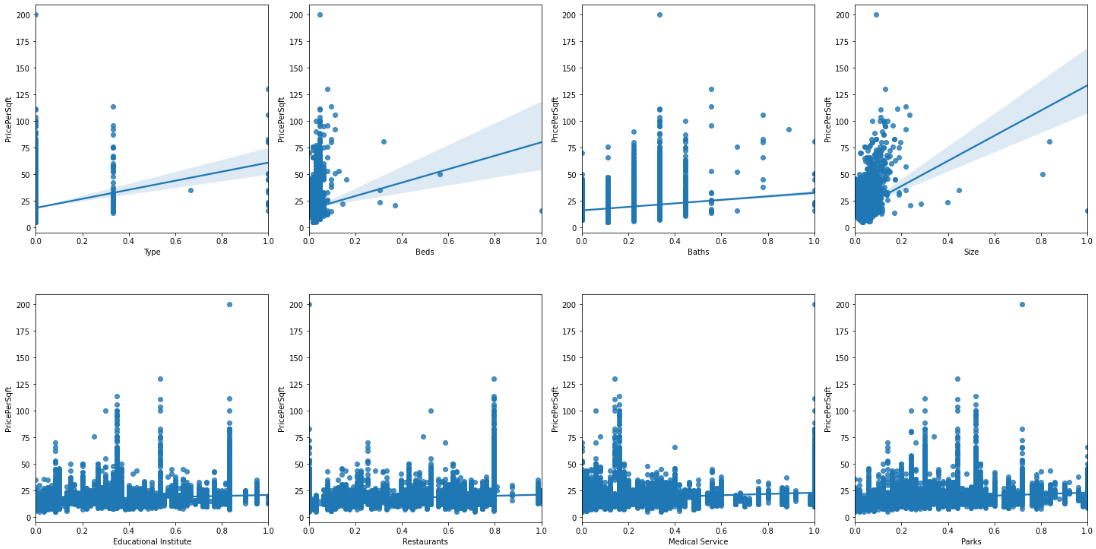
Figure 10 depicts the distribution of all traits in relation to density.

Figure 11,12: PricePerSqft and Price density visualizing using Distribution plot



Figures 11 and 12 exhibit the PricePerSqft and Price against the density, respectively. The distribution of PricePerSqft is more evenly distributed in this case, but the distribution of Price is not. Perhaps this is the cause for the regression model's lower accuracy. PricePerSqft is a better goal value than price, according to this distribution plot graph.

Figure 13 : Visualize all the features against PricePersqft using Reg plot



Here in figure 13 all the features are plotted in a Reg plot against the target value which is PricePersqft.

F. Data Cleaning

Another important task is data cleansing. When cleaning data, it's examined to see whether there are any missing values. When training and testing the regression model, missing data might lead to unexpected outcomes. 'Beds', 'Baths', 'Educational Institute', 'Restaurants', 'Medical Service', and 'Parks' all had missing data initially. After that, all of the missing values were enumerated, and then all of the missing values were restored using the interpolation approach by estimating the neighbor value. After discovering a large number of duplicate values in the data, all duplicate values were replaced, and all entities with missing values were eliminated.

Table 4: Comparing missing value before and after removing duplicate values

Before removing duplicate value		**Before removing duplicate value**	
Address	False	Address	False
Type	False	Type	False
Beds	True	Beds	False
Baths	True	Baths	False
Size	False	Size	False
Educational Institute	True	Educational Institute	False
Restaurants	True	Restaurants	False
Medical Service	True	Medical Service	False
Parks	True	Parks	False
PricePerSqft	False	PricePerSqft	False

G. Features Selection

The process of limiting the number of input variables is known as feature selection. The majority of the time, all of the input variables are ineffective in constructing a strong predictive model. Dropping some of the input variables may result in a decent prediction model in this scenario. In most cases, if a feature has a negative correlation with the target values, it will not perform well in the model. As a result, removing the characteristics is beneficial to the model.

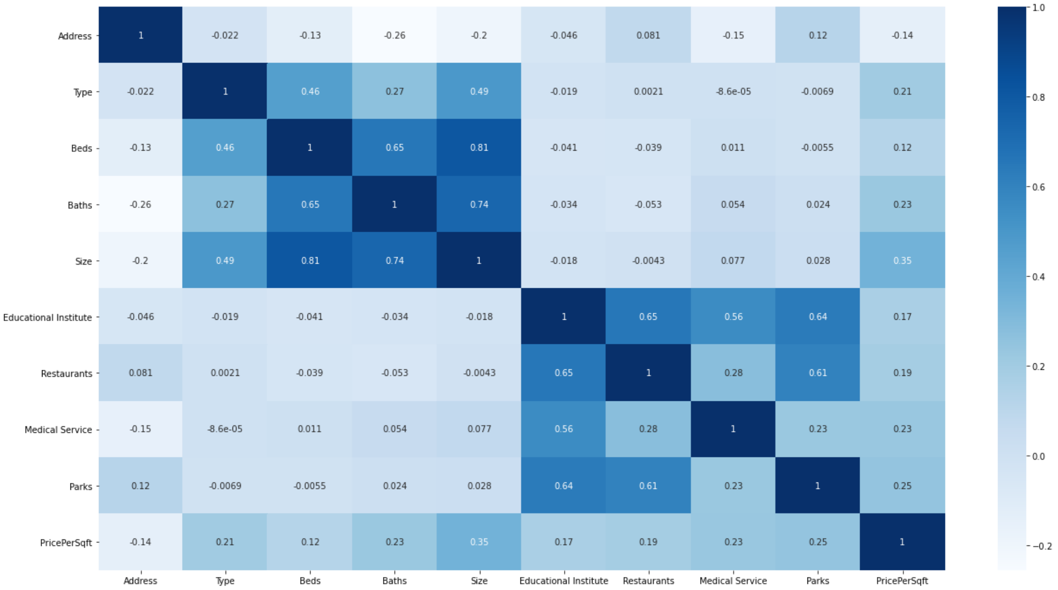
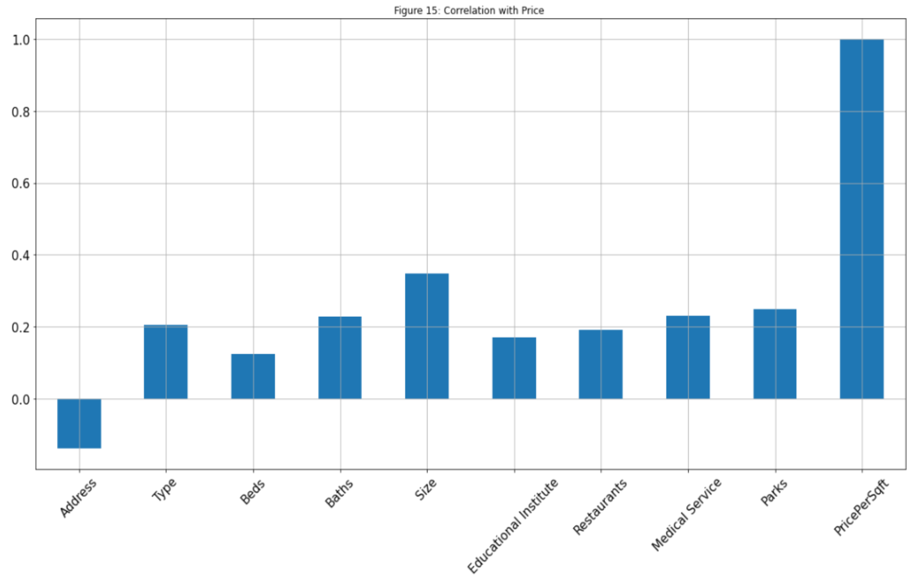


Figure 14: Heatmap

A heatmap based on the association of each variable with each variable is displayed in figure 14. It's crucial to pay attention to the last column or row in this case. The association with the goal value PricePersqft is shown in the last portion and column. Only the variable 'Address' shows a negative association with the goal value in this case. Figure 15 plots all of the factors that are related to the price per square for a better understanding. The 'Address' characteristic shows an obvious negative link with the price per square foot. As a result, removing 'Address' may result in improved performance.



H. Data Split

The data that will be fed into the model is in the train set, whereas the data that will be used to validate and test the trained model is in the test set. A test dataset and a training dataset can be prepared in a variety of ways. It will also work fine with two separate datasets for train and test. However, the core data set is split into two parts: a training dataset and a test dataset. The test size is set to 0.3, which implies that the entire dataset will be divided into three parts, two of which will be used for training and the remaining portion will be utilized for testing.

I. Data Scaling

The range of raw data values vary substantially in some machine learning algorithms, and these widely variable data do not function effectively without normalization or standardization. To deal with the widely variable pricing per square foot, the data were normalized and standardized for this project. Later on, during the model train, all three data were used to get three separate assessment results: original, normalized, and standardized values.

J. Regressors Used

A total of six regression methods were employed in this research. They are linear regression, ridge regression, lasso regression, decision tree, Random Forest, and extreme gradient boosting.

Linear Regression:

Linear regression[6] is a regression model where the features (X) and the target class (y) variables are considered to have a linear relationship. The target class is determined by the linear combination of the input variables in this model. It can be a univariate linear regression or a multivariate linear regression, depending on the number of variables.

A linear regression line's equation is as follows:

$$\hat{y} = a + bX$$

The predicted dependent variable is $\hat{y}$,, and the independent variable is X. The line's slope is b, and the intercept is a.

Ridge Regression:

Ridge regression[7] is a model tuning technique for analyzing multicollinear data. L2 regularization is used in this procedure. When there is a problem with multicollinearity, the least-squares method is unbiased, and the variances are enormous, resulting in projected values that are distant from the actual values.

The cost function for ridge regression:

$$\boldsymbol{Min(||Y - X(\theta)||\hat{}2 + \lambda||(\theta)||\hat{}2)}$$

LASSO Regression:

LASSO[8] is the short form of Least Absolute Shrinkage and Selection Operator. It is a shrinkage-based linear regression model. The central point where the data is shrunk like the mean is called shrinkage. This regression is also known as the L1 regularize, and it is especially useful when there are fewer parameters and a high degree of multicollinearity. The purpose of LASSO regression is to reduce the cost factors as much as possible.

The equation for a linear regression line is as follows:

$$Cost(\beta) = \sum_{i=1}^{n} (y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Here, λ denotes the amount of shrinkage.

Decision Tree:

A supervised machine learning approach called Decision Tree[9] is used to create regression or classification models. It utilizes a tree-like structure, with the leaf nodes representing the outcomes. Except for the leaf nodes, all other nodes are considered decision nodes, where more splits are made based on yes/no questions. The goal of the decision tree is to construct a model that can predict the value of a target variable by learning basic choice rules from past data. The entropy or Gini index is frequently used to determine how a decision tree splits the data.

Random Forest:

Random Forest[10] is an ensemble learning approach that creates a classification or regression model by combining numerous decision trees. A random forest is made up of large numbers of decision trees. that operate together as a group. Individual trees predict the target class's value, and their predictions are aggregated to provide a more accurate prediction. Random forest is a supervised machine learning approach for solving classification and regression problems.

Extreme Gradient Boosting:

Extreme Gradient[11] Boosting is based on the Gradient Boosting technique, a machine learning approach for solving classification and regression problems. XGBoost implements the gradient boosting technique in an efficient and effective manner. It improves performance by combining a number of weak prediction models. Multiple decision trees are commonly utilized. To control overfitting in severe gradient boosting, a more regularized model is adopted, which increases its performance even further.

K. Model Evaluation

RMSE:

The standard deviation of the residuals or prediction errors is known as the Root Mean Square Error (RMSE). The residuals are a measure of how distant the data points are from the regression line. The RMSE is a measure of how evenly distributed the residuals are. In other words, it indicates how tightly the data is clustered around the line of best fit. In climatology, forecasting, and regression analysis, root mean square error is widely used to check experimental results.

MSE:

The mean squared error (MSE) measures the distance between a regression line and a set of points. It accomplishes it by squaring the distances between the points and the regression line. Squaring is required to eliminate any negative signals. Because we're calculating the average of a set of errors, it's termed the mean squared error.

MAE:

The magnitude of the difference between an observation's predicted value and its real value is referred to as absolute error. The total magnitude of the group's errors is measured by the average of absolute errors for a group of predictions and observations. MAE, being one of the most widely used loss functions for regression problems, assists the user in transforming learning problems into optimization problems. It also functions as a basic, quantifiable measurement of errors for regression problems.

R2 Score:

The R2 score, also known as the coefficient of determination, is used to evaluate the efficacy of a linear regression model. The degree of variance in the output dependent characteristic can be predicted based on the input independent variable (s).

Cross Validation:

Cross-validation is a method of training our model using a subset of the data set and then evaluating it using the other subset. That is, to use a small sample to evaluate how the model will perform in general when used to make predictions on data that was not utilized during the model's training.

IV. Results

Table 5: Regression algorithm evaluation with different data scale

Regression Algorithm	Evaluation Method	RMSE	MSE	MAE	R2 Score	Cross Validation
	Data Scale					
Linear Regression	Original	6.1127	37.3662	4.1125	0.2600	0.2228
	Normalized	6.1127	37.3662	4.1125	0.2600	0.2228
	Standardized	6.1127	37.3662	4.1125	0.2600	0.2228
Ridge regression	Original	8.5692	73.4317	3.8039	-0.4541	0.2228
	Normalized	8.5692	73.4317	3.8039	-0.4541	0.2228
	Standardized	8.5692	73.4317	3.8039	-0.4541	0.2228
LASSO Regression	Original	7.6500	58.5228	3.7852	-0.1588	0.3100
	Normalized	7.6500	58.5228	3.7852	-0.1588	0.3100
	Standardized	7.6500	58.5228	3.7852	-0.1588	0.3100
Decision Tree	Original	5.1776	26.8076	3.4223	0.4691	0.3331
	Normalized	5.1829	26.8633	3.4241	0.4680	0.3331
	Standardized	5.1785	26.8175	3.4233	0.4689	0.3331
Random Forest	Original	4.6488	21.6119	3.0789	0.5720	0.5089
	Normalized	4.6524	21.6448	3.0811	0.5713	0.5089
	Standardized	4.6506	21.6287	3.0806	0.5717	0.5089
XG Boosting	Original	8.5778	73.5791	6.4151	-0.4570	0.5085
	Normalized	8.5778	73.5791	6.4151	-0.4570	0.5089
	Standardized	8.5778	73.5791	6.4151	-0.4570	0.5089

During the model train, original, normalized, and standardized values data were used to provide three separate assessment findings. The RMSC, MSC, MAE, R2 Score, and Cross Validation techniques were used to evaluate all six regression algorithms. Smaller numbers are preferable in RMSC, MSC, and MAE assessing techniques. In the case of R2 Score and Cross Validation, however, the higher the score, the better.

In the table for all the regression algorithms original, normalized and standardized values are showing the same performance in case of decision tree and random forest the value is little bit changed in case of three different data scales. This difference can be neglected. So, here data scaling has no effect on all these regression models.

The lesser the Root Mean Square Error (RMSE), the better the performance. The method with the greatest RMSC value is XG boosting and ridge regression, while the approach with the lowest score is Random Forest Regression with a score of 4.6488. As a result, the RMSC score indicates that Random Forest Regression will perform better. It is similar to the RMSE in the case of Mean Squared Error (MSE). In this examination, XG boosting and ridge regression have the greatest score, whereas random forest has only a score of 21.6287. Random Forest is superior. Now, according to MAE, XG boosting has a value of 6.4151, whereas Random Forest has a value of 3.0789, indicating that Random Forest is the best. Now we'll look at the R2 score and Cross Validation. The better the performance, the higher the score. The R2 scores for XG Boosting, LASSO Regression, and Ridge Regression were all negative, whereas Random Forest had the maximum positive value of 0.5720. Without a doubt, the R2 score indicates that Random Forest is superior. Only in Cross Validation does XG Boosting perform better than Random Forest regression. XG Boosting, on the other hand, does poorly in other evaluation models.

After reviewing all of the scores, it is evident that the Random Forest Regressor outperforms all other assessment techniques, and it is best to avoid using XG Boosting for prediction due to its poor performance.

V. Conclusions

This research analyzes a variety of machine learning approaches to predict the rental price of housing by comparing different characteristics. The dataset for this research was obtained from the website BProperty.com. We found out that type, *beds, baths, educational institutes, restaurants, medical services, and parks* are the most important features for this prediction by eliminating outliers and irrelevant features from the dataset. After experimenting with a range of machine learning algorithms on our gathered data, we discover that Random Forest outperforms all model evaluation methods, including RMSE, MSE, MAE, R2 Score, and Cross Validation.

Reference

[1] Neloy, A. A., Haque, H. M. S., & Ul Islam, M. M., (22 February 2019) “Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring.” Proceedings of the 2019 11th International Conference on Machine Learning and Computing - ICMLC ’19. [Online] Available: https://dl.acm.org/doi/abs/10.1145/3318299.3318377

[2] Yue Ming , Jie Zhang , Jiaming Qi , Tian Liao , Maolin Wang and Lingli Zhang, (23 October 2020) “Prediction and Analysis of Chengdu Housing Rent Based on XGBoost Algorithm,” ICBDT 2020: Proceedings of the 2020 3rd International Conference on Big Data Technologies. [Online] Available: https://dl.acm.org/doi/abs/10.1145/3422713.3422720

[3] Embaye WT, Zereyesus YA and Chen B, (11 February 2021) “Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches,” PLoS ONE 16(2): e0244953. [Online] Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0244953

[4] Takahiro Yoshida and Hajime Seya, (27 July 2021) “Spatial prediction of apartment rent using regression based and machine learning-based approaches with a large dataset,” arXiv preprint. [Online] Available: https://arxiv.org/abs/2107.12539

[5] F Yue and W Yingnian, (2020) “California Rental Price Prediction Using Machine Learning Algorithms,” University of California, Los Angeles. [Online] Available: https://escholarship.org/uc/item/0h04h8ms

[6] Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining, (April 2012) “Introduction to Linear Regression Analysis, 5th Edition”

[7] Ashok, P. (2020, 10 15). *Great Learning*. Retrieved from What is Ridge Regression? [Online] Available: https://www.mygreatlearning.com/blog/what-is-ridge-regression/?fbclid=IwAR2H6wluV9RFlnwBkr5HGJtCE_B_kPGsWXFu6JDNLTJd86MQc_dylOjyRUs

[8] Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J Roy Statist Soc ser B 58(1):267–288

[9] Quinlan, J. R. (1986). Induction of decision trees . *Springer Nature* . https://link.springer.com/chapter/10.1007%2F978-981-16-1220-6_29

[10] Andy Liaw and Matthew Wiener, (December 2002) “Classification and Regression by RandomForest,” [Online] Available:

https://www.researchgate.net/publication/228451484_Classification_and_Regression_by_RandomForest

[11] Jerome H.Friedman, (28 February 2002) “Stochastic gradient boosting,” Department of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94305, USA. [Online] Available: https://www.sciencedirect.com/science/article/abs/pii/S0167947301000652?via%3Dihub