# Real Estate Company Beater Price Prediction Using Machine Learning

Supervisor: Professor Dr. Shazzad Hosain
Md. Nur-E-Azam –1512268042
Md. Reaz Islam – 1520663042
Hridoy Saha - 1611520042

Department of Electrical and Computer Engineering (ECE) , North South University (NSU), Bashundhara, Dhaka-1229, Bangladesh

Email: nur.azam@northsouth.edu

*Is it possible to predict real estate house predictions effectively using Machine learning algorithms and advanced data mining tools.*

**Abstract**

The following paper covers the implementation of the price prediction project for the housing and immobilization markets. Many algorithms are used to improve the percentage effectively, numerous researchers have conducted this project and are using algorithms such as hedonic regression and artificial neural networks that are regarded the best pricing models.. They are regarded by advanced data analytics algorithms such as random forests, gradient trees, multi-layer perceptrons and ensemble machine learning models, using these models and achieving prediction accuracy at a greater rate.

**Keywords:** Random Forest, DecisionTreeRegressor,Linear Regression, Gradient Boosting Regressor,Multi layer perceptron,price prediction, Python, Sklearn, Pandas, NumPy,Machine learning,Advanced ML Algorithm and Model.

## Introduction

We need a predictive approach to the real estate and houses on the housing market, but there are many who are making big mistakes in Bangladesh and South Asian countries right now when they buy property most people are buying properties without being seen a mechanism to buy and sell a house for most of the people. In 2017 there was an estimate of around 5.42 million homes sold in the country of South Asia in 2016 but of 10.7 percent below 2015. Therefore, there was a reduction of the initial home stock.

In 2007 and 2008 an economic collapse occurred so that several economic indicators indicated the imminent disaster, and currently this situation is occurring and economic indicators suggest that high housing prices are becoming known by high people who use the real estate to know the economic situation in force.

In general, property may have to give the worth of the land. Many diverse

players in the commercial center carry out a quantitative measure of profit, such as land agents, assessors, mortarboards, brokers, developers, gurus Book managers, loan companies, etc. Through this request, business value will be evaluated. From claiming value systems Methods reflecting the kind and condition of property offered by them.

The property may well under various conditions and circumstances be exchanged on the open market, the persons are unaware of the current conditions and begin to lose their money. In order to prevent some circumstances the changes in property prices could affect the ordinary population as well as the government.

In this I try to anticipate the forecast of the future real estate price using the machine learning techniques utilizing the past works, many methods have been utilized as a hedonic regression. I utilized the random forest, the Tree Regressor and other algorithms to estimate property prices with various tools It would thus be useful for people to be aware of both present and future conditions so that they can prevent errors. The rest of the paper is arranged as Section 2 summarizes the past work done by several researchers using various algorithms The technique and tools utilized are provided in section 3, and section 4 covers the implementation of the algorithms, and last part provides for comparisons and outcomes.

## Related Work

A literature review must be done before committing to the project many basic works I reviewed various documents on property markets and other areas relevant to the price forecast. Our major aim is to get greater precise results than prior efforts, in the various years to the current year and I've applied the latest technology. Past predictive work by diverse scholars is described in the following sections and the appropriate project will be beneficial to execute.

## Summary of Literature Review

The literary review gives a clearly defined notion for each project, and serves as the basis here for most writers to conclude that artificial neural networks have a more effective influence on the prediction but that other algorithms must also be taken into account in the actual world. It enabled me to know both pros and maize by performing this study and had helped me implement this project successfully.

## Methodology

The next sections detail the approach employed in the projections of property prices and the flow diagram for architecture is presented.

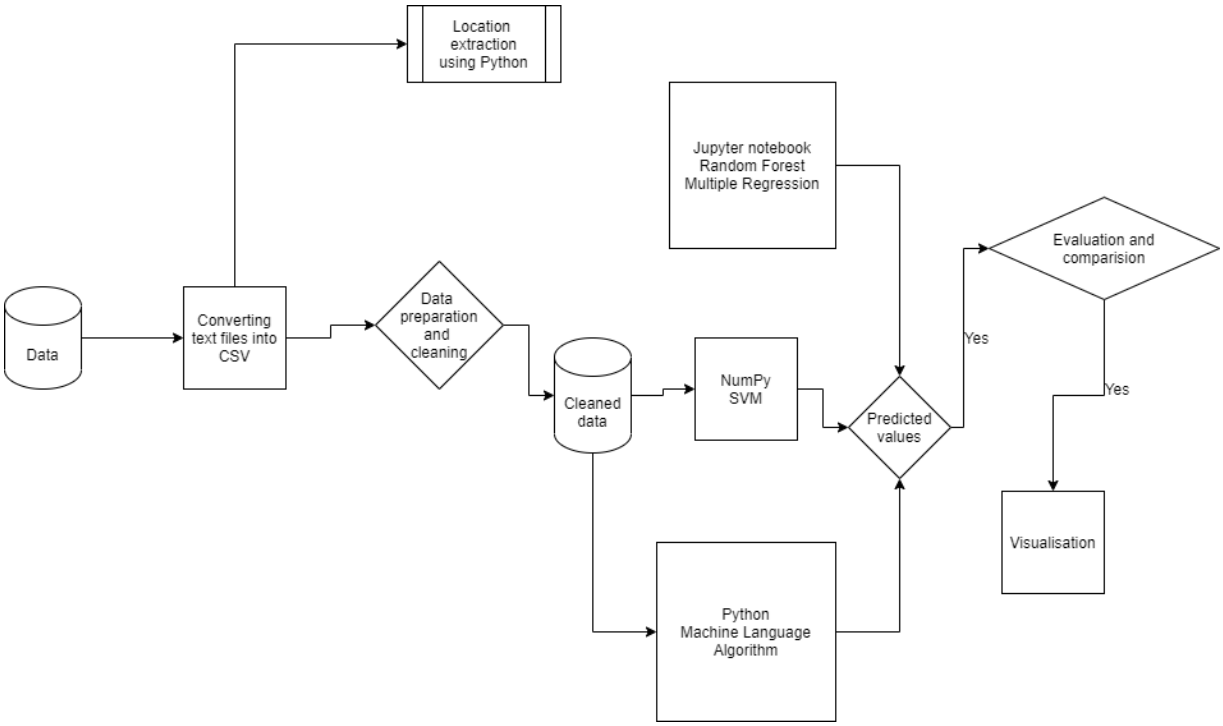


Figure 2: Architecture of price prediction

**Description of Data-sets**

The real estate housing data is used in this and it is taken from the UCI machine learning repository and the ageron the data is spread across 506 rows and has the thirteen attributes and 1 binary-valued attribute the description of the data set is given below

| **S.no** | **Variables** | No. of Instances |
|---|---|---|
| 1 | CRIM | 506 |
| 2 | ZN | 506 |
| 3 | INDUS | 506 |
| 4 | CHAS | 506 |
| 5 | NOX | 506 |
| 6 | RM | 506 |
| 7 | AGE | 506 |
| 8 | DIS | 506 |
| 9 | RAD | 506 |
| 10 | TAX | 506 |
| 11 | PTRATIO | 506 |
| 12 | B | 506 |
| 13 | LSTAT | 506 |
| 14 | MEDV [ Special Attribute] | 506 [value to be predicted] |

Here there are totally 13 predictor variables and the MEDV variable will be median house price which is going to be predicted.

### Data Cleaning And integration

The data from the repository is in the text file I linked to the Excel and extracted the data from the text file, moved to the Excel file and saved as a comma-separating file. Data purification is an iterative process, the first iteration is to find and rectify misinformation. There are many inconsistencies in the repository, and zero values in data taken from the repository before being loaded into machine learning models should be corrected in order to achieve high prediction accuracy, because the different prediction tools use, and the process of cleaning differs from one other. The real estate data have some missing information were given by using the Python program. I have identified and take care of missing value, set the value to some value(0, mean or median) and the null values are removed to reduce the inconsistency.

### Detection Of Outliers

In information can be determined whether the outlier is an exceptionally high or very low number, whether the value is more than the interquartile range. A data order is arranged between the bottom value and the greater value in Q3 + 1.5 or Q1 - 1.5, the average value is now taken for the first set of values and second set values, by subtracting the two average interquartile format for the interquartile range Q3 + (1.5) for the first set of values and the Q1-(1.5) for the second set values, and I calculated with the Python software.

### Tools

Here there are some essential tools used for the prediction project.

| Tools and Algorithms | | |
|---|---|---|
| S.no | Name of the tools | Algorithms used |
| 1<br>2<br>3 | Jupyter notebook<br>Microsoft Excel sheet<br>Anaconda Python | Random forset, Linear Regression, Multiple Regression, Decision Tree, Gradient Boosting Trees, Machine Learning, Bagging |

### Regression

It is a data analysis task of predicting the value of target(numerical variable) by building a model based on the one or more predictors the predictors can either be numerical or the categorical variables.

### Machine Learning Algorithms

It may be used to predict both the classification and regression as the regression forests. Random forest technique may be employed. The fundamental procedure is the creation of a large random data selection, random variables and dependent variable classes based on many different trees. The principal advantage of utilizing this dataset methodology is that it handles the missing values and can keep the missing data accurate and the chances of overriding the model is minimal. When we apply to large-level datasets, we are except highly sized. The result will be continuous in regression trees.

### Multiple Regression

It is a new version of a linear regression that is regarded as more powerful and works

with multiple variables and multiple features to predict an unknown value of the feature from the known value of the two or more predictors.

### Decision Tree

A Decision Tree is a fluctuation chart structure with a test of an attribute on an internal node (e.g. if a coin flip comes up with heads or tails), each branch reflects the test result and each board node represents a class marker (decision taken after computing all attributes).

### Gradient Boosting

Gradient boosting may be utilized both for regression and classification, as indicated by (Ganjisaffar et al.;2011) Gradient boosting is a technique for constructing regression models made up of regressor sets This notion for regression is an instantiation The essential issue is to follow the technique again and again Here we learn the simple data regression forecast, then we calculate the rest of the mistake. The error quantity per data point and we learn to anticipate the residual error in a new model. The key notion is that we make a series of forecasts to detect and reduce the inaccuracies.

### Machine Learning

Machine learning is a way to analyze data that automates the construction of analytical model. It has the premise that systems can learn from data, find designs and decide by minimizing human involvement. It is an artificial intelligence industry.

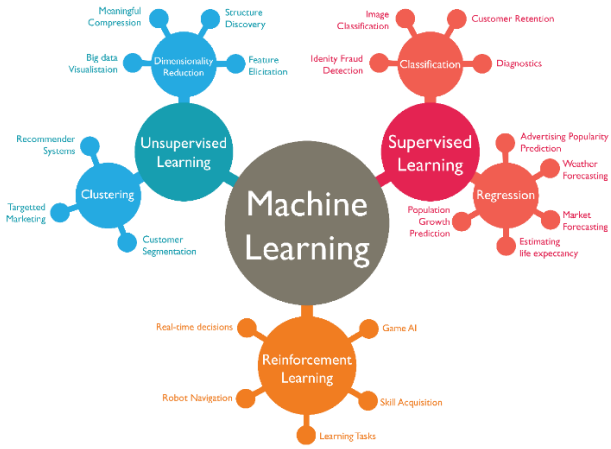


Figure 3: Machine Learning

### Ensemble learning Bagging

It is part of a machine learning technique known as a group learning method, aiming to increase stability and minimize variation and accuracy. The use of group learning is nothing more than that, so that effectively numerous models are created and a more accurate model is developed. When storing different models, different samples are constructed simultaneously to each other and then the different models vote for the final model and consequent forecasting.

### Accuracy Calculation And visualisation

The metrics I'm using to assess performance correctness are the mean absolute error, the difference between the forecast and the actual number. The findings are loaded into the table after prediction so that they may be readily shown and utilized for future activities as well.

### Implementation

The main aim of this project to be implemented is to find out the accurate prediction of the real-estate properties present in Bangadeshe for the next upcoming years, the below segment blankets will help you to know the implementation process in depth. Here step

by step process involved is represented below.

1. Scientific Environment.
2. Source of a Data.
3. Excel 2016: the first process to store the data.
4. Loading data into Python, Excel, Jupyter Notebook.
5. Normalizing the data.
6. Detecting Outliers.
7. Analysis and visualization using the Python, Jupyter Notebook, Anaconda Python.
8. Machine learning models are build using the Cat Tools, and the various algorithm used for predictions as listed in the methodologies.
9. Splitting the data sets as test and train for the Cross-validation process.
10. Fitting the data into different machine learning models and Algorithm for the Predictions.
11. Finding the Root mean square value and MSE mean squared error to finding the Accuracy percentage.
12. Visualization.

**Need of a Technical Environment**

1. Microsoft Excel
2. Jupyter Notebook for creating the Scripts
3. Anaconda
4. Pandas
5. Appropriate Functions to be selected in the Python programming language
6. Matplotlib For visualization and plotting histogram

**Data Source**

As I said before in the key data sets, this is an open datasource for data mining and predictive analytics from the UCI machine learning repository. The data source obtained was a text file that identifies data source problems, connects the text file to the EXCEL and separates it using commas and is stored as a CSV file.

**Data cleaning**

The data in the CSV files must be examined for missed values since each attribute examined using the filters has missing data source, while null values are eliminated to raise the degree of accuracy.

**Implementation of Random Forset and Multiple Regression**

**Packages Used:** Random Forest, numpy, sklearn

The data is been split into test and train 80 percent data is used for the train data and remaining for the test data since. I am using the Random forest the number of trees used is 200.Consider the standard recursive partitioning algorithm will start it searches all the data and in-depth search is made for all the variables and the best-split point is taken and the process gets repeated for the right and left leaves. Here in the random forest, the variables are selected randomly from our variables with the help of predictor attributes for each split the different variables are selected. we can find the variable importance, so here the median house value has the more influence in predicting. The prices were predicted and the graph shows. Here the average output of several trees has been taken in order to provide more accuracy.

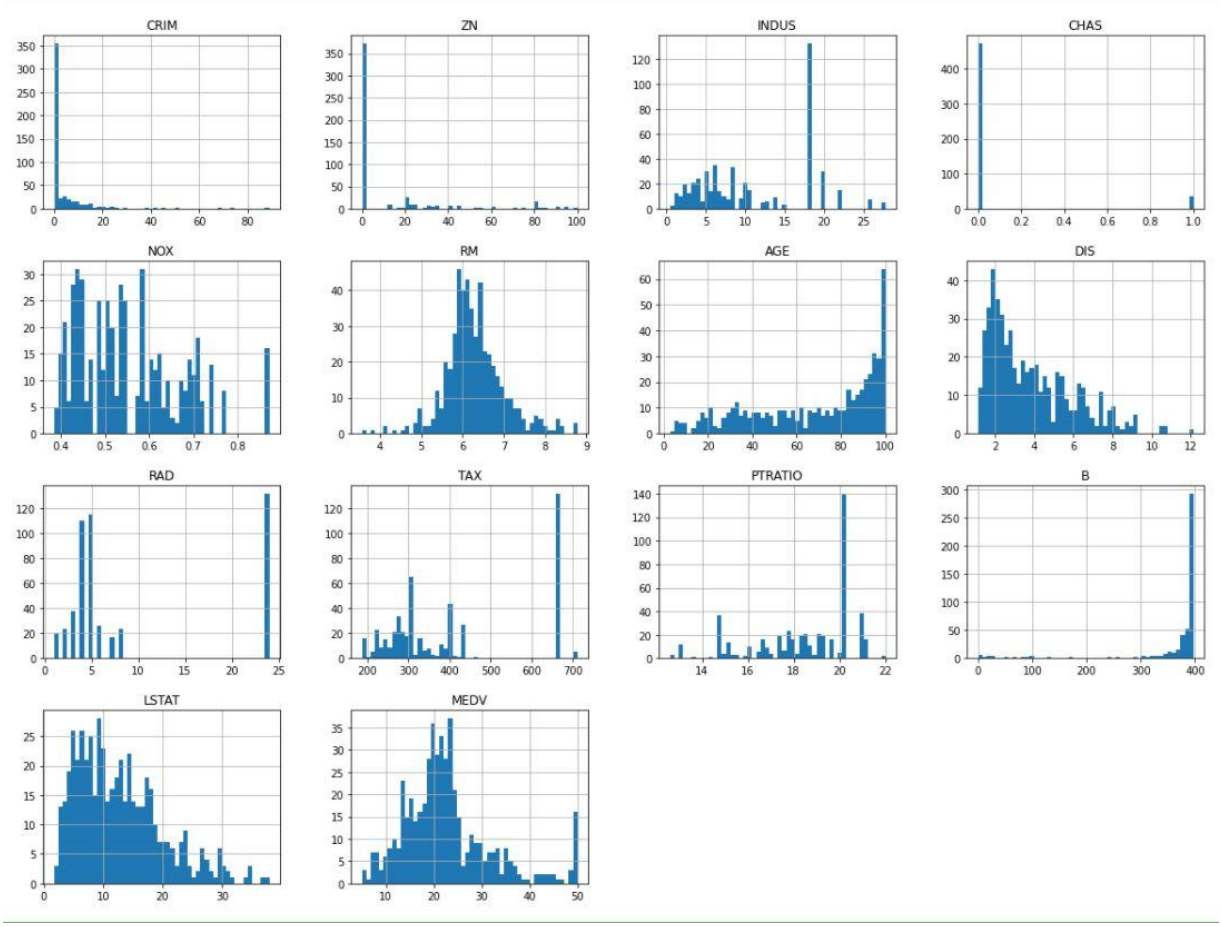


Figure 4: feature extraction(variable importance)

The same procedure was used with many regression assuming the dependent variable is a median house value attribute to be expected and the remaining characteristics are independent variables. The characteristics are indicated in the model and an equation for regression is generated with the help of coefficients.

## Implementation Of SVM and Gradient Boosted

Python algorithms in Anaconda are pre-determined when the loaded data is related to a function called the selected attributes here where the not null values are assumed to avoid the null values in the prediction process, each attribute has only a single role to play in predicting the attribute using a set role function.The cross validation is performed in two sections one, and the following tests the data set. The fundamental principle of cross validation is to divide the data into k bins of identical size such that the data is segregated here according to containers I use validation of 10 fold crosses Nine parts of data are being used to test data and the other part is being used as the algorithms we need to train the data, I have used the gradient boosted tree and the vector support machine. The test section may both anticipate the data and classify the performance. The vector count performance support aids to generate the predictable variables and the performance regression will aid in the RMSE mean square error values, which effectively

anticipate prices and discuss error proportions in the evaluation part.

### Evaluation

The goal is to forecast an actual value that implies that number may be calculated in the regression model The following are the most prevalent terms:

### Coefficient of determination

The coefficient of determination mse summarizes the explanatory power of the regression model and is computed from the sum of squares terms. The mse describes the proportion of variance of the dependent variable explained by the regression model and the equation is given below

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 5: mse Formula

The figure shows the correlation between my attributes and how they are related with each other

MEDV 1.000000
RM 0.680857
B 0.361761
ZN 0.339741
DIS 0.240451
CHAS 0.205066
AGE -0.364596
RAD -0.374693
CRIM -0.393715
NOX -0.422873
TAX -0.456657
INDUS -0.473516
PTRATIO -0.493534
LSTAT -0.740494
Name: MEDV, dtype: float64

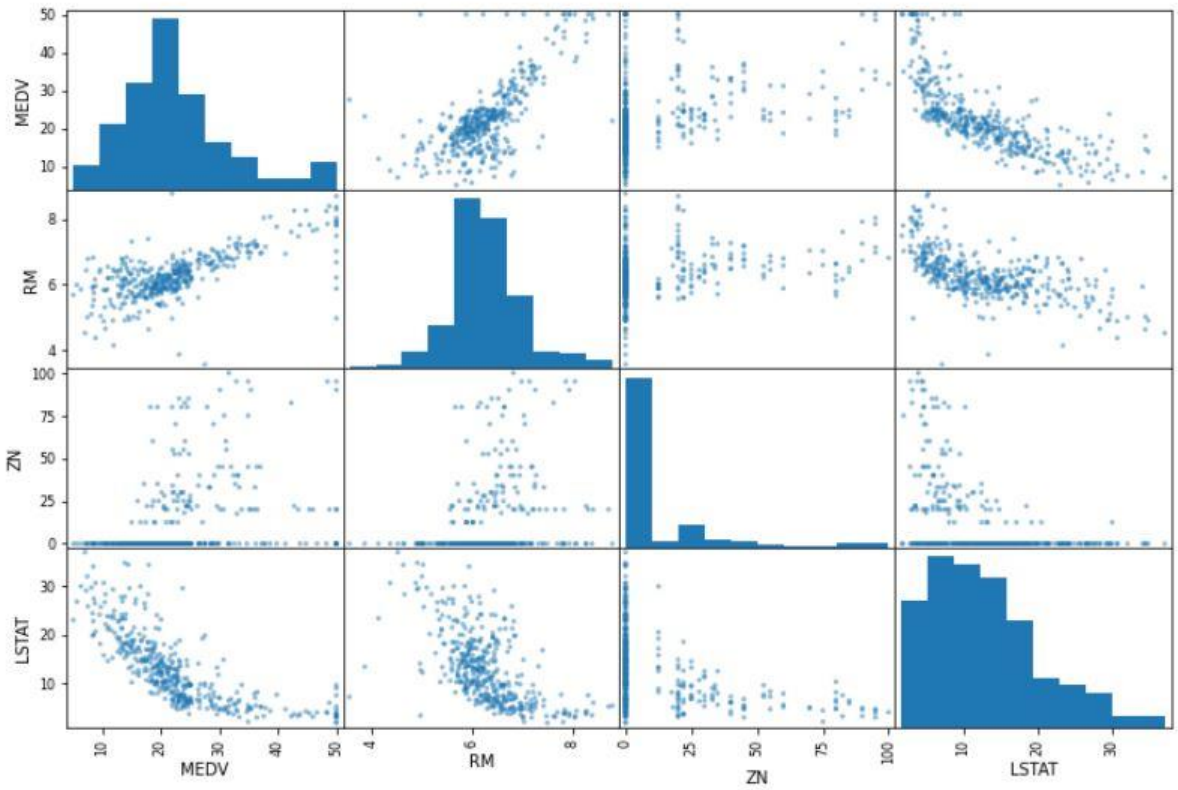


Figure 6: Correlation between attribu

**Root Mean Square Error**

RMSE is a popular formula for measuring a regression model's error rate, however, only models may be compared with errors evaluated in the same units using the supplied formula.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

Figure 7: rmse formula

Where N is the data number of instances, x is the projected I-instance value and x'^ is the real value The fundamental notion is to subtract from the actual value square the expected value, that the RMSE is accomplished by obtaining the total of all instances and dividing it by the number of cases. The main factors used to compute the error value are the following: the following table summarizes the root mean squares of the different algorithms, and helps you ascertain how effectively the Algorithm can forecast future values.

**Evaluation of Result**

| RMSE |
|---|

| S.no | Algorithms | RMSE Error | Accuracy |
|---|---|---|---|
| 1 | Random Forest Regressor | 0.1875517 | 95% |
| 2 | Linear Regression | 0.8353010 | 50% |
| 3 | Decision Tree Regressor | 0.0 (overfitting) | Overfitting |
| 4 | Gradient Boosting | 0.3649096 | 75% |
| 5 | Extra Tree Regressor | 0.0(overfitting) | Overfitting |
| 6 | Multiple Regression | 0.7043517 | 58% |
| 7 | Support Vector machine | 0.6365109 | 60% |

**Discussion**

This experiment with diverse machine learning algorithms shows that a random set-up and gradient-boosted stress work better with a higher precision % and lower error values. These algorithms forecast well when comparing this experiment to the result produced. Machine learning and associated technology and algorithm have been carried out in this project.

## Conclusion

The principal purpose of the project is to determine price forecasts that we succeeded in using several master learning algorithms, such a Random Forest, multiple regression, vector support machines, gradient boosted trees, the Decision Bree Regressor and an Extra Tree Regressor. I thus feel that this research will aid both the public and the governments and the works below.

Each system and new technology in the software can assist estimate costs in future. This may be enhanced by incorporating a wide range of factors, such as the environment, the markets and other associated housing characteristics. You may save the expected information in databases and make an app for customers so that they have a short concept and invest money in a more secure method. When there is the potential of real-time data, the data can be connected to H2O and machine learning algorithms can be connected directly to the interlink.


## Acknowledgements

I would thank my supervisor Professor Dr. Shazzad Hosain (SZZ) for providing his time in guiding me and the knowledge he shared with me, his guidance is the main reason to complete and successfully implement this project.

# References


B.V., E. (2020, 1 20). *sciencedirect*. Retrieved from Housing Price Prediction via Improved Machine Learning Techniques: https://www.sciencedirect.com/science/article/pii/S1877050920316318

Bhagat, N. M. (2016). House price forecasting using data mining. *International Journal of Computer Applications*, 152(2): 23–26.

K., P. B. (2015). Using machine learning algorithms for housing price. *The case of fairfax county, virginia housing data, Expert Systems with*, 2928–2934.

Kabir, R. (2013, 4 23). *brainstation-23*. Retrieved from Real Estate House Price Prediction: Model Building: https://brainstation-23.com/a-real-life-use-case-of-machine-learning-technology/

Khan, Y. (2020, sep 4). *Using Machine Learning Algorithms to Predict Pricing Trends*. Retrieved from onetech.ai: https://www.onetech.ai/en/blog/using-machine-learning-algorithms-to-predict-pricing-trends#:~:text=With%20Machine%20Learning%20(ML)%20technology,the%20target%20variable%20is%20numeric.

Liaw, A. W. (2002). Classification and regression by randomforest. *Classification and regression by randomforest, R news*, 2(3): 18–22.

Limsombuncha. (2004). House price prediction: hedonic price model vs. artificial. *New Zealand Agricultural and Resource Economics Society Conference*, pp. 25–26.

Murage, A. ( 2020, November 23). *House Price Prediction using Machine Learning*. Retrieved from section.io/engineering-education: https://www.section.io/engineering-education/house-price-prediction/

Piazzesi, M. a. (2009). Momentum traders in the housing market. *survey evidence and a search model*, 101-115.

Ravikumar. (2018). House Price Prediction Using Machine Learning and Neural Networks. *House Price Prediction Using Machine Learning and Neural Networks*, 18-21.

Selim, H. (2009). Determinants of house prices in turkey. *Hedonic regression versus artificial neural network*, 2843–2852.

Willmott, C. J. (1981). On the validation of models. *On the validation of models*, 2(2): 184–194.