A Project ETE Report

on

REAL ESTATE PRICE PREDICTION

Submitted in partial fulfillment of
the requirement for the award of
the degree of

B.TECH CSE - AIML

Under The
Supervision of
Mr. S. Kalidass
ASST. PROFESSOR

Submitted By

RISHIV KAUSHAL 19021130031
SARVESH SHARMA 19021180090

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING GALGOTIAS UNIVERSITY, GREATER NOIDA,

INDIA
12,2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"REAL ESTATE PRICE PREDICTION'** in partial fulfillment of the requirements for the award of the **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** submitted in

the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of JULY,2021 to DECEMBER,2021 under the supervision of Mr. S. KALIDASS, Assistant Professor, Department of Computer Science and Engineering of School of Computing Science and Engineering, Galgotias University, Greater Noida

The matter presented in the project has not been submitted by us for the award of any other degree of this or any other places.

RISHIV KAUSHAL - 19SCSE1180124

SARVESH SHARMA – 19SCSE1180100

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Mr. S KALIDASS
ASSISTANT PROFESSOR

CERTIFICATE

The Final Project Viva-Voce examination of RISHIV KAUSHAL: 19SCSE1180124 AND SARVESH SHARMA: 19SCSE1180100 has been held on_____and his work is recommended for the award of BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: DECEMBER,2021

Place: Greater Noida

ABSTRACT

There are several objects that are often used for investment, for example, gold, stocks and property. In particular, property investment has increased significantly. Housing price trends are not only the concern of buyers and sellers, but it also indicates the current economic situation. There are many factors which has impact on house prices, such as location, BHK, floor etc. Also, a location with a great accessibility to highways, expressways, schools, shopping malls and local employment opportunities contributes to the rise in house price.
Manual house prediction becomes difficult, hence there are many systems developed for house price prediction.

Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of our project. Investment is a business activity on which most people are interested in this globalization era. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price. We use various regression techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied. In this we will use the boston dataset. This pipeline creation process involves loading the dataset, cleaning and pre-processing the dataset, fitting a model to the dataset, and testing the model's performance using various evaluation metrics

STUDENT TABLE:

S.No	Name	Enrollment No.	Admission No.	Program /Branch	Sem	Mobile	Email
1.	Rishiv Kaushal	19021130031	19SCSE1180124	Btech CSE – AIML	5	7906178271	Rishiv_kaushal.scsebtech@galgotiassuniversity.edu.in
2.	Sarvesh Sharma	19021180090	19SCSE1180100	Btech CSE - AIML	5	8295848290	Sarvesh_sharma.scsebtech@gAlgotiasuniversity.edu.in

Faculty Details:

Name	Designation	Email id	Mobile No.
Mr. S. Kalidass	Assistant Professor	s.kalidass@galgotiasuniversity.edu.in	8939673979

List of Figures

ACRONYMS

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering
SVM	Support Vector Machine
RCNN	Region Based Convolutional Neural Networks
FFT	Fast Fourier Transformation
SVD	Singular Valued Decomposition
PCA	Principal Component Analysis
WEKA	Waikato Environment for Knowledge Analysis
BiRNN	Bidirectional Recurrent Neural Networks

Table Of Contents

INTRODUCTION

The study on land price trend is deemed to be significant to support the decisions in urban planning. The real estate system is an unstable stochastic process. Investors decisions are based on the market trends to reap maximum returns. Developers are interested to know the future trends for their decision making. In order to accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis, modelling and forecasting. The factors that affect the land price have to be studied and their impact on price has also to be modelled. An analysis of the past data revealed that the prices show a non- linear characteristic. It is inferred that establishing a simple linear mathematical relationship for these time-series data is found not viable for forecasting. Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyze and forecast future trends.

As the real estate is fast developing sector, the analysis and forecast of land prices using mathematical modelling and other scientific techniques is an immediate urgent need for decision making by all those concerned. The increase in population as well as the industrial activity is attributed to various factors, the most prominent being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started of showing an upward trend and housing and the real estate activity started booming. All barren lands and paddy fields ceased their existence to pave way for multistore and high-rise buildings. Investments in Real Estate Industry has grown significantly high over the years and we have noticed a non-uniform pattern in terms of land pricing. The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors.

So, we are going to make an end to end ML model by using different types of regression algorithms to predict the price of the property and then choose the suitable algorithm which gives less error and more accurate value.

TOOLS AND FRAMEWORK

1) **Scikit learn -** (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.
2) **Matplotlib -** Matplotlib is an amazing visualization library in Python for 2D plots of arrays. it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.
3) **Pandas -** It is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.
4) **NumPy -** NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

5) **Seaborn** - Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis.

6) **Google Colab -** Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

LITERATURE SURVEY

The latest worldwide financial crisis restored a sharp enthusiasm toward both academic and strategy circles on the part of asset costs and specifically lodging costs clinched alongside monetary movement. As Lamer (2007) notes those lodging showcase predicted eight of the ten post globe War ii recessions, acting Concerning illustration An heading woman for those true segment of the economy. Truth be told he dives Likewise significantly Concerning illustration with state that “Housing is those benefits of the business cycle”.

Recently, a few writers scope to experimental discoveries that house costs can make instrumental molding to determining yield.

There is huge literature writing in regards to U.S. house prices. Rapach Furthermore strauss (2007) use an auto regressive dispersed slack (ARDL) model framework, holding 25 determinants with conjecture genuine lodging cost development to the unique states of the elected Reserve’s eighth region. They discover that ARDL models tend should beat a benchmark AR model.

Numerous relapse Investigation will be very nearly the same Likewise basic straight relapse. The main distinction the middle of straightforward straight relapse Also numerous relapse is in the number for predictors (“x” variables) utilized within those relapse.

Straightforward relapse examination employments An absolute x variable to each subordinate “y” variable. Case in point: (x1, Y1). Numerous relapse utilization numerous “x” variables for every free variable: (x1)1, (x2)1,(x3)1, Y1). In one-variable straight regression, you might information particular case subordinate variable (i. E. “sales”) against a autonomous variable (i. E. “profit”). Anyhow you could make intrigued by how diverse sorts from claiming offers impact the relapse. You Might set your X1 as particular case kind from claiming sales, your X2 Similarly as in turn sort about deals etc.

In the twenty-first century, real estate has evolved into much more than a necessity; it now represents much more. Not limited to those looking to buy Real Estate, but also for those who sell it. Real estate, according to [4], is not only a man's basic need, but it also represents a person's wealth and prestige. Because property values do not decline overnight, real estate investment appears to be profitable in general. changes in real estate prices can have an impact on a various types of people, including homeowners, bankers, policymakers, and others. Property investment appears to be an attractive option for investors. As a result, predicting real estate value is an important economic indicator. This tells us that every organisation in today's

real estate business is working hard to gain a competitive advantage over other competitors. The process must be simplified for the average person while still producing the best results.[6] proposed employing ML and AI techniques to develop an algorithm capable of predicting property prices based on specific input features. This algorithm has a business application in that known websites can use it directly to forecast the prices of new properties that will be listed by some inputting variables and forecasting the correct and justified price, i.e., avoid taking price inputs from clients and therefore not allowing any error to creep into the software. the Google Colab/Jupiter IDE Jupiter IDE is a free and open-source web application that lets us share and create documents with LiveCode, visualisations, equations, and narrative text. It includes data cleaning, data transformation, numerical value simulation, statistical modelling, data visualisation, and machine learning tools. developed a system to assist people in determining the real estate price The user can enter their specifications and receive the prices of the desired houses. A sample plan of the house can also be obtained by the user as a reference for houses. In, the SVM, LSSVM, and PLS methods, as well as the corresponding characteristics, are used to analyse and forecast the housing value of a Boston suburb. After removing the missing data from the original data set, some samples are classified as training data and rest as test data. The housing value of the training data. The Random Forest Regressor algorithm outperformed the Decision Tree Regressor algorithm in terms of accuracy. Ridge and Linear Regression produce comparable results, with the exception of a very slight reduction in Lasso. It is a good indication that the purchasing prices can be used solely to predict the selling prices without taking other attributes into consideration to reduce model to overfit. Moreover, accuracy reduces in the very weak features group. For all feature selections, the Root Square Mean Error (RMSE) exhibits the same pattern of results. Another study found that preparing their data set took more than a day. Instead of performing the computations sequentially, we could use multiple processors to parallelize the computations, potentially reducing the preparation and prediction time. In addition to the model, we can give customers the option of selecting a city or a locality to produce high-temperature maps, rather than including them in the list. It made use of a data set of 100 houses with various parameters. We trained the machine with half of the data set and tested it with the other half. The results are 100 % correct. We also managed to run it through a set of tests with different parameters. Regression is used because not using PSO makes it a lot simpler to train machines with difficult problems. tested the most basic valueDecision tree classifiers, decision tree regression, and multiple linear regression are examples of machine learning algorithms. To accomplish the task, the Scikit-Learn machine learning library is used. This system facilitates users in

predicting the availability of houses in the city as well as their own prices. Machine learning algorithms have been used to forecast house prices. We walked you through the process of examining the dataset step by step. These feature sets were then used into four algorithms, which resulted in a csv file with predicted housing prices. it stated that a combination of these models is required because a linear model produces a high bias (underfit), whereas a high model complexity-based model produces a high variance (overfit). These findings can be used in the annual updation of the guideline value of land, potentially adding more revenue to the State Government when the transaction is completed. By running the model with different machine learning algorithms, it is clear that random forest tress and gradient boosting trees perform better in terms of accuracy percentage and error values. These techniques predict well when the results of this research are compared to the label.

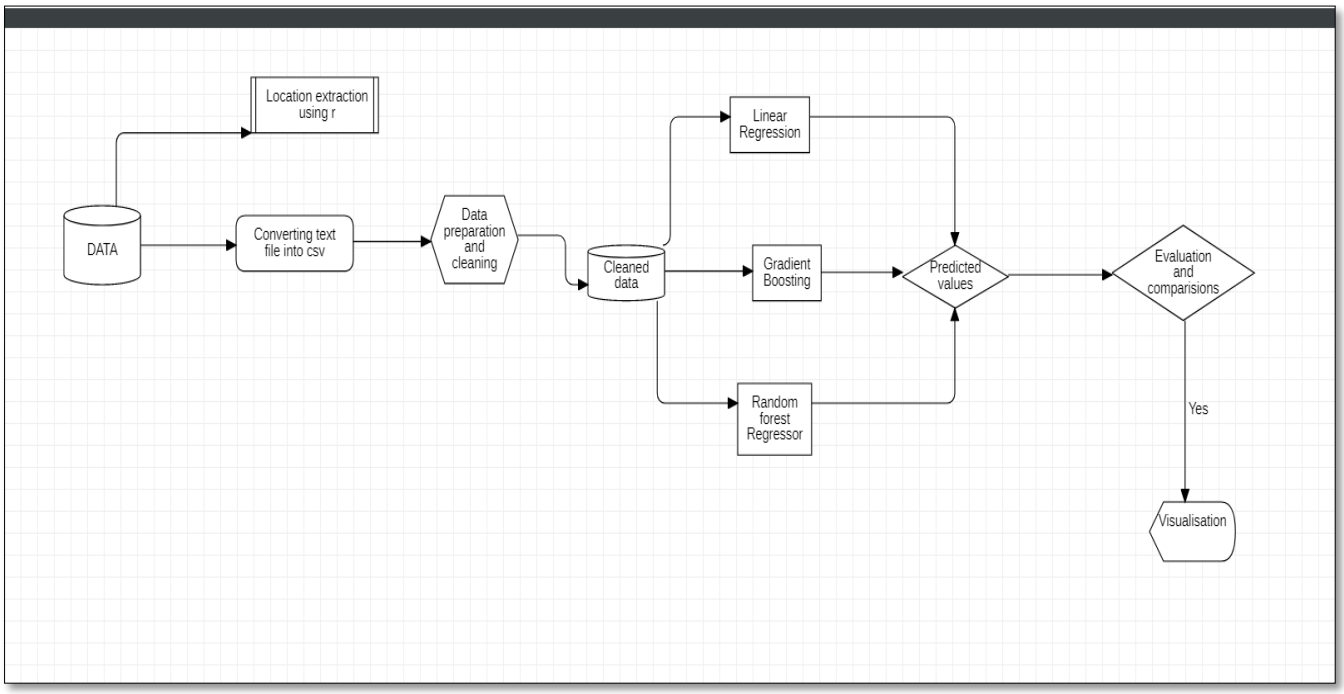
PROPOSED MODEL

We are going to make an end to end ML model to predict the prices of real estate. In our model the user has to input the numerical values of the features used in our model and our model will predict the price of the desired property. In this model we are going to use various machine learning algorithms then observe the results from different algorithms and choose which gives the most accurate price. In our model we currently plan using two datasets which are Boston housing dataset and Mumbai pricing dataset.

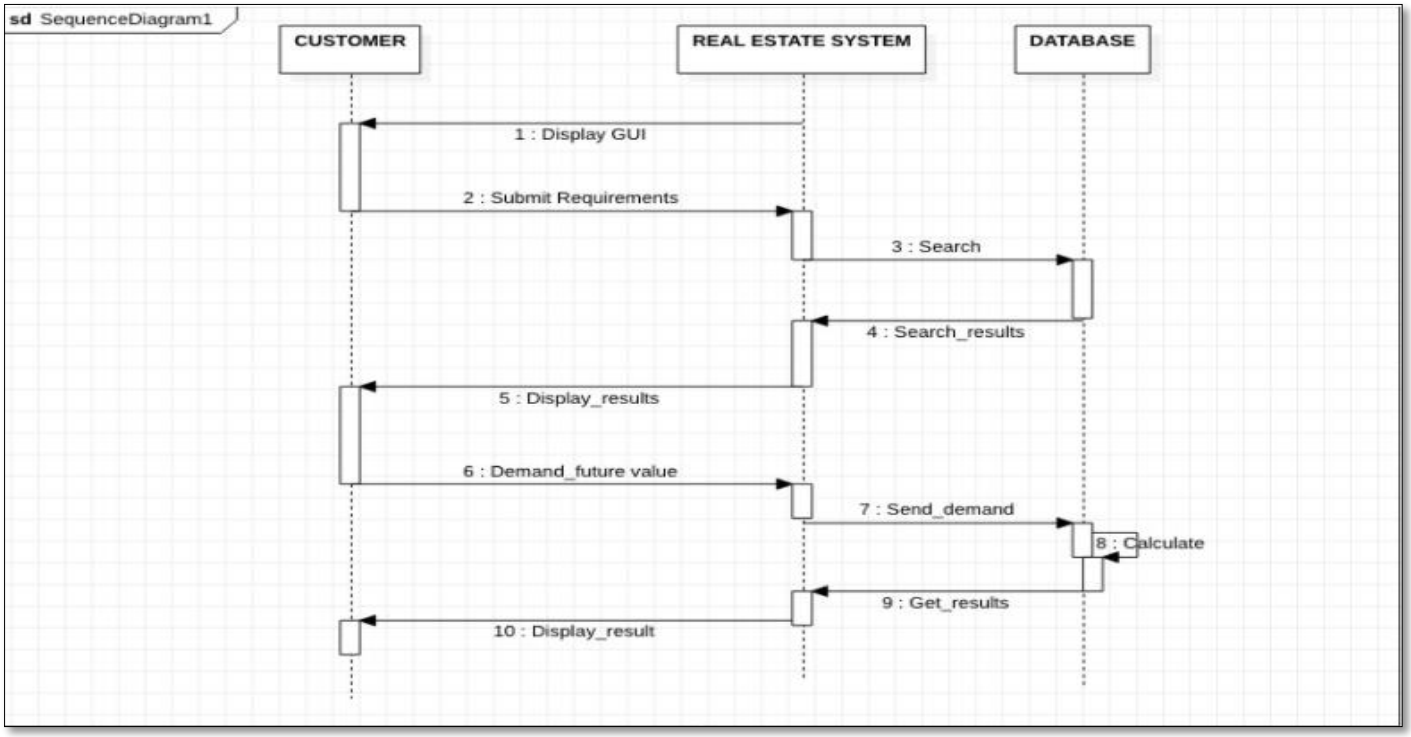
ROADMAP OF THE MODEL

1) At first, we are going to take the dataset from the sources like KAGGLE, BOSTON UCI, etc.
2) Then we will convert the dataset into csv file
3) The we will clean the data i.e. we will if there is any data missing or repeated.
4) Then we will apply the different types of machine learning algorithms on the cleaned data.
5) After applying the algorithms, we will analyze the results from different algorithms and compare it.
6) In the end, it will forecast the most accurate value.

DESIGN OF THE MODEL



SEQUENCE DIAGRAM



DATASET

HOUSING PRICES IN MUMBAI

This data contains scraped data and has the following information:

1) Prices of houses all over Mumbai along with their location
2) Information about house condition (new/resale) and area of the house
3) Information about various amenities provided

FEATURES

1) Area- Size of the house in square feet.
2) Location in which area of Mumbai.
3) No. of Bedrooms
4) Resale
5) Gymnasium
6) Swimming pool
7) Landscaped garden
8) Jogging track
9) Shopping mall
10) ATM
11) Clubhouse
12) School
13) 24X7 security
14) Power backup
15) Car parking
16) Hospital
17) Gas connection
18) Lift available

ALGORITHMS USED

1. LINEAR REGRESSION

Linear regression is an attractive model because the representation is so simple.
The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

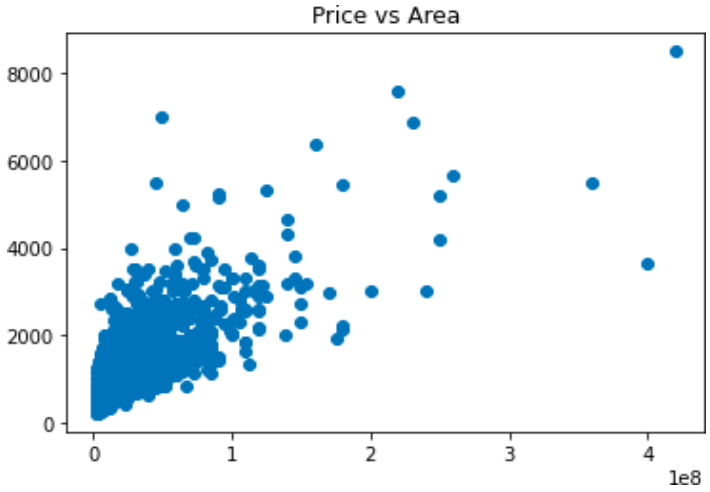
For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

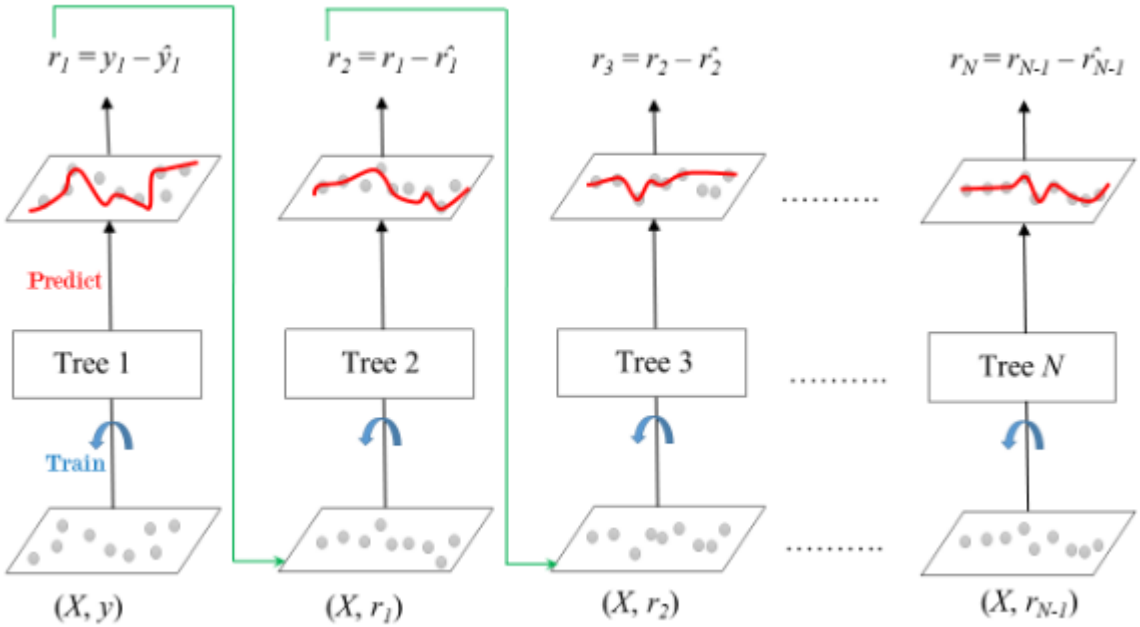
It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model ($0 * x = 0$). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.



2. GRADIENT BOOSTING ALGORITHM

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Ada boost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. There is a technique called the **Gradient Boosted Trees** whose base learner is CART (Classification and Regression Trees).



The ensemble consists of *N* trees. Tree1 is trained using the feature matrix *X* and the labels *y*. The predictions labelled *y1(hat)* are used to determine the training set residual errors *r1*. Tree2 is then trained using the feature matrix *X* and the residual errors *r1* of Tree1 as labels. The predicted results *r1(hat)* are then used to determine the residual *r2*. The process is repeated until all the *N* trees forming the ensemble are trained. There is an important parameter used in this technique known as Shrinkage.

Shrinkage refers to the fact that the prediction of each tree in the ensemble is shrunk after it is multiplied by the learning rate (eta) which ranges between 0 to 1. There is a trade-off between eta and number of estimators, decreasing learning rate needs to be compensated with increasing estimators in order to reach certain model performance. Since all trees are trained now, predictions can be made.

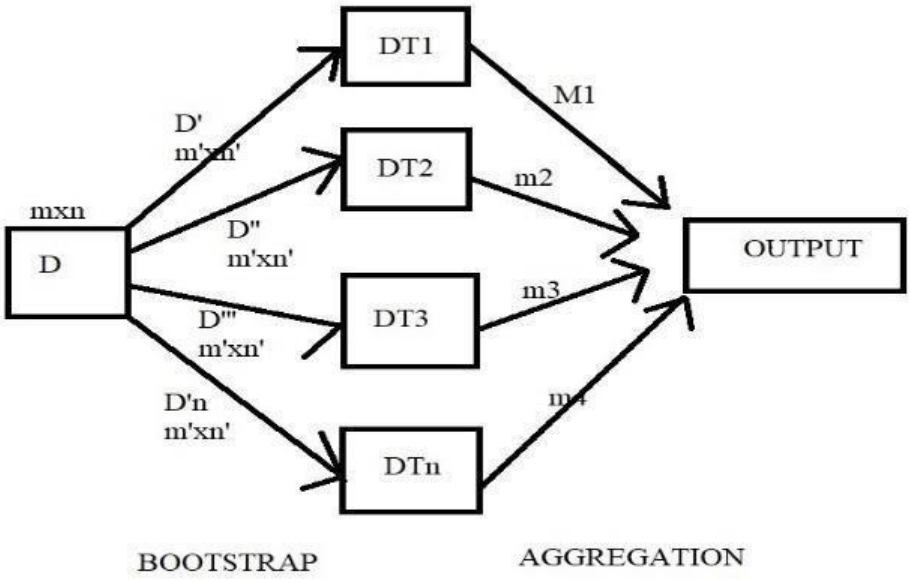
3. RANDOM FOREST REGRESSION

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. It generates predictions without requiring many configurations in packages (like scikit-learn).



How random forest algorithm works

Understanding decision trees

Decision trees are the building blocks of a random forest algorithm. A decision tree is a decision support technique that forms a tree-like structure. An overview of decision trees will help us understand how random forest algorithms work.

A decision tree consists of three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further segregate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be segregated further.

The nodes in the decision tree represent attributes that are used for predicting the outcome. Decision nodes provide a link to the leaves. The following diagram shows the three types of nodes in a decision tree.

The information theory can provide more information on how decision trees work. Entropy and information gain are the building blocks of decision trees. An overview of these fundamental concepts will improve our understanding of how decision trees are built.

Entropy is a metric for calculating uncertainty. Information gain is a measure of how uncertainty in the target variable is reduced, given a set of independent variables.

The information gain concept involves using independent variables (features) to gain information about a target variable (class). The entropy of the target variable (Y) and the conditional entropy of Y (given X) are used to estimate the information gain. In this case, the conditional entropy is subtracted from the entropy of Y.

Information gain is used in the training of decision trees. It helps in reducing uncertainty in these trees. A high information gain means that a high degree of uncertainty (information entropy) has been removed. Entropy and information gain are important in splitting branches, which is an important activity in the construction of decision trees.

Let's take a simple example of how a decision tree works. Suppose we want to predict if a customer will purchase a mobile phone or not. The features of the phone form the basis of his decision. This analysis can be presented in a decision tree diagram.

The root node and decision nodes of the decision represent the features of the phone mentioned above. The leaf node represents the final output, either *buying* or *not buying*. The main features that determine the choice include the price, internal storage, and Random Access Memory (RAM).

Applying decision trees in random forest

The main difference between the decision tree algorithm and the random forest algorithm is that establishing root nodes and segregating nodes is done randomly in the latter. The random forest employs the bagging method to generate the required prediction.

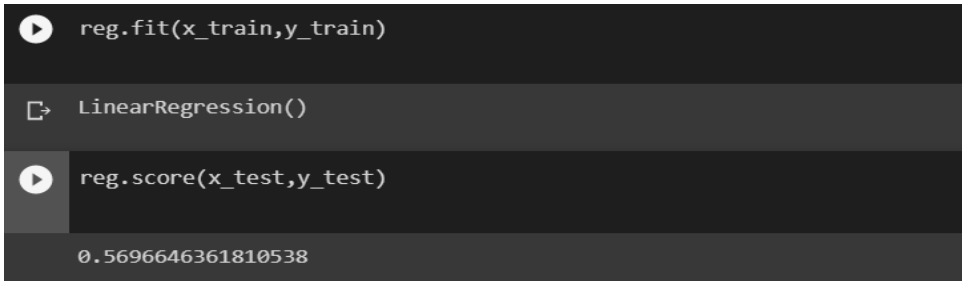
Bagging involves using different samples of data (training data) rather than just one sample. A training dataset comprises observations and features that are used for making predictions. The decision trees produce different outputs, depending on the training data fed to the random forest algorithm. These outputs will be ranked, and the highest will be selected as the final output.

Our first example can still be used to explain how random forests work. Instead of having a single decision tree, the random forest will have many decision trees. Let's assume we have only four decision trees. In this case, the training data comprising the phone's observations and features will be divided into four root nodes.

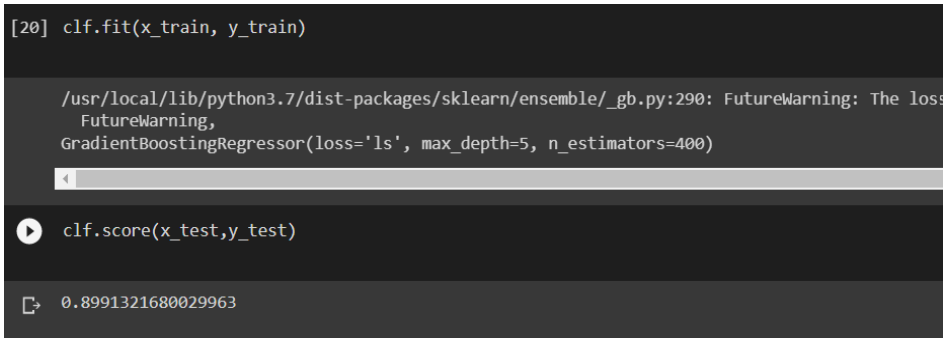
The root nodes could represent four features that could influence the customer's choice (price, internal storage, camera, and RAM). The random forest will split the nodes by selecting features randomly. The final prediction will be selected based on the outcome of the four trees.

RESULTS

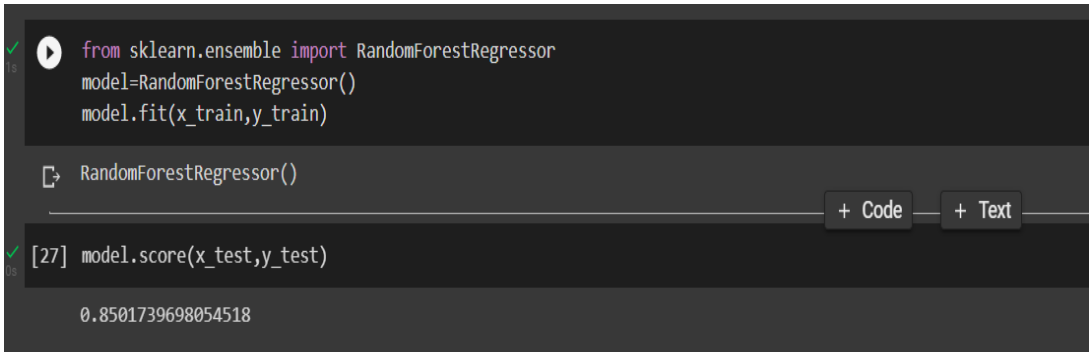
1. LINEAR REGRESSION



2. GRADIENT BOOSTING ALGORITHM



3. RANDOM FOREST REGRESSION



CONCLUSION

The system is capable of training itself and predicting prices based on the raw data provided to it. After reviewing several research papers as well as numerous blogs and articles, a set of algorithms that were suitable for use on both datasets of the model were chosen. Following multiple testing and training sessions, it was determined that the XGBoost Algorithm outperformed the other algorithms. The system was powerful enough to predict the prices of various houses with various features, and it could handle large amounts of data. The system is simple to use and saves time.

A supplementary feature that can be added to our proposed system is to provide users with a full-fledged user interface, so that users can use the ML model for multiple locations with multiple functionalities. A connection to Amazon EC2 will also take the system a step further and make it easier to use. Finally, the project will be completed by creating a well-integrated web application that can predict prices whenever users want it to.

REFERENCES

1. Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*, 18–22.
2. Limsombunchai, V. (2004). House price prediction: hedonic price model vs. artificial neural network. In *New Zealand Agricultural and Resource Economics Society Conference* (pp. 25–26).
3. Nghiep, N., & Al, C. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research*, *22*(3), 313–336G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu , "House Price Prediction Using Machine Learning" , International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-9, 2019.
4. Kuvalekar, Alisha and Manchewar, Shivani and Mahadik, Sidhika and Jawale, Shila, House Price Forecasting Using Machine Learning (April 8, 2020). Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) 2020
5. Neelam Shinde, Kiran Gawande , "Valuation Of House Prices Using Predictive Techniques", International Journal of Advances in Electronics and Computer Science, Volume-5, Issue-6, 2018.
6. Jingyi Mu, Fang Wu,and Aihua Zhang , " Housing Value Forecasting Based on Machine Learning Methods", Hindawi Publishing Corporation Abstract and Applied Analysis, Volume 2014.