International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

Real Estate Price Prediction

Smith Dabreo[1], Shaleel Rodrigues[2], Valiant Rodrigues[3], Parshvi Shah[4]
[1, 2, 3] Student, Fr. Conceicao Rodrigues College of Engineering, Mumbai.
[4]Assistant Professor, Fr. Conceicao Rodrigues College of Engineering, Mumbai.

Abstract:- **This paper demonstrates the usage of machine learning algorithms in the prediction of Real estate/House prices on two real datasets downloaded from Kaggle from Boston created by Harrison, D., and Rubinfeld, D.L. and from Melbourne created by Anthony Pino. To this day, literature about research on machine learning prediction of house prices in India is extremely limited. This paper provides a review of the usage of existing machine learning algorithms on two extremely different datasets and tries to implement this prediction engine for real-life usage by users. The findings indicate that using different algorithms can drastically change accuracy. Also, a poor dataset can negatively affect the predictions. Furthermore, it provides sufficient proof of what algorithm is best suitable for this task.**

Keywords: Machine Learning, Real Estate, House Price, Price Prediction, Algorithm.

I. INTRODUCTION

Machine Learning (ML) is a vital aspect of present-day business and research. It progressively improves the performance of computer systems by using algorithms and neural network models. Machine Learning algorithms automatically build a mathematical model using sample data – also referred to as "training data" which form decisions without being specifically programmed to make those decisions.
People and real estate agencies buy or sell houses, people buy to live in or as an investment and the agencies buy to run a business. Either way, we believe everyone should get exactly what they pay for. over-valuation/under-valuation in housing markets has always been an issue and there is a lack of proper detection measures. Broad measures, like house/Real-estate price-to-rent ratios, give a primary pass. However, to decide about this issue an in-depth analysis and judgment are necessary. Here's where machine learning comes in, by training an ML model with hundreds and thousands of data a solution can be developed which will be powerful enough to predict prices accurately and can cater to everyone's needs.

The primary aim of this paper is to use these Machine Learning Techniques and curate them into ML models which can then serve the users. The main objective of a Buyer is to search for their dream house which has all the amenities they need. Furthermore, they look for these houses/Real estates with a price in mind and there is no guarantee that they will get the product for a deserving price and not overpriced. Similarly, A seller looks for a certain number that they can put on the estate as a price tag and this cannot be just a wild guess, lots of research needs to be put to conclude a valuation of a house. Additionally, there exists a possibility of underpricing the product. If the price is predicted for these users this might help them get estates for their deserving prices not more not less.

II. LIMITATIONS OF PREVAILING METHODOLOGIES

There is a notable amount of research done in the house price prediction department but very research has come up to any real-life solutions. There is very little evidence of a working house price predictor set up by a company. For now, very few digital solutions exist for such a huge market and most of the methods used by people and companies are as follows:

Buyers/Customers:

1. When people first think of buying a house/Real estate they tend to go online and try to study trends and other related stuff. People do this so they can look for a house which contains everything they need. While doing these people make a note of the price which goes with these houses. However, the average person doesn't have detailed knowledge and accurate information about what the actual price should be. This can lead to misinformation as they believe the prices mentioned on the internet to be authentic.
2. The second thing that comes to mind while searching for a property is to contact various Estate agents. The problem with this is these agents need to be paid a fraction of the amount just for searching a house and setting a price tag for you. In most cases, this price tag is blindly believed by people because they have no other options. There might be cases that the agents and sellers may have a secret dealing and the customer might be sold an overpriced house without his/her knowledge.

Seller/Agencies:

1. When an individual thinks of selling his/her property they compare their property with hundreds and thousands of other properties which are posted all around the world. Determining the price by comparing it with multiple estates is highly time-consuming and has a potential risk of incorrect pricing.
2. Large Real estate companies have various products they need to sell and they have to assign people to handle each of these products. This again bases the prediction of a price tag on a human hence there is room for human error. Additionally, these assigned individuals need to be paid. However, having a computer do this work for you by crunching the heavy numbers can save a lot of time money and provide accuracy which a human cannot achieve.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

III. LITERATURE REVIEW

Real Estate has become more than a necessity in this 21st century, it represents something much more nowadays. Not only for people looking into buying Real Estate but also the companies that sell these Estates. According to [4] Real Estate Property is not only the basic need of a man but today it also represents the riches and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Changes in the real estate price can affect various household investors, bankers, policymakers, and many. Investment in the real estate sector seems to be an attractive choice for investments. Thus, predicting the real estate value is an important economic index. [3] suggests that Every single organization in today's real estate business is operating fruitfully to achieve a competitive edge over alternative competitors. There is a need to simplify the process for a normal human being while providing the best results. [6] proposed to use machine learning and artificial intelligence techniques to develop an algorithm that can predict housing prices based on certain input features. The business application of this algorithm is that classified websites can directly use this algorithm to predict prices of new properties that are going to be listed by taking some input variables and predicting the correct and justified price i.e., avoid taking price inputs from customers and thus not letting any error creeping in the system. [12] used Google Colab/Jupiter IDE. Jupiter IDE is an open-source web app that helps us to share as well create documents that have LiveCode, visualizations, equations, and text that narrates. It contains tools for data cleaning, data transformation, simulation of numeric values, modeling using statistics, visualization of data, and machine learning tools. [10] designed a system that will help people to know close to the precise price of real estate. User can give their requirements according to which they will get the prices of the desired houses User can also get the sample plan of the house to get a reference for houses. In [5] Housing value of the Boston suburb is analyzed and forecast by SVM, LSSVM, and PLS methods and the corresponding characteristics. After getting rid of the missing samples from the original data set, 400 samples are treated as training data and 52 samples are treated as test data. Housing value of the training data. As per [1]'s findings, the best accuracy was provided by the Random Forest Regressor followed by the Decision Tree Regressor. A similar result is generated by the Ridge and Linear Regression with a very slight reduction in Lasso. Across all groups of feature selections, there is no extreme difference between all regardless of strong or weak groups. It gives a good sign that the buying prices can be solely used for predicting the selling prices without considering other features to disseminate model over-fitting. Additionally, a reduction in accuracy is apparent in the very weak features group. The same pattern of results is visible on the Root Square Mean Error (RMSE) for all feature selections. [2] observed that their data set took more than one day to prepare. As opposed to performing the computations sequentially, we might utilize various processors and parallel the computations involved, which might possibly decrease the preparation time Furthermore prediction period. Include All the more functionalities under the model, we can give choices for clients with select a district alternately locale should produce those high-temperature maps, as opposed to entering in the list. [7] used a data set of 100 houses with several parameters. We have used 50 percent of the data set to train the machine and 50 percent to test the machine. The results are truly accurate. And we have tested it with different parameters also. Not using PSO makes it easier to train machines with complex problems and hence regression is used. [13] experimented with the most fundamental machine learning algorithms like decision tree classifier, decision tree regression, and multiple linear regression. Work is implemented using the Scikit-Learn machine learning tool. This work helps the users to predict the availability of houses in the city and also to predict the prices of the houses. [8] used machine learning algorithms to predict house prices. We have mentioned the step-by-step procedure to analyze the dataset. These feature sets were then given as an input to four algorithms and a CSV file was generated consisting of predicted house prices. [9] expressed that There is a need to use a mix of these models a linear model gives a high bias (underfit) whereas a high model complexity-based model gives a high variance (overfit). The outcome of this study can be used in the annual revision of the guideline value of land which may add more revenue to the State Government while this transaction is made. [11] concludes that by conducting this experiment with various machine learning algorithms it's been clear that random forest and gradient boosted trees are performing better with more accuracy percentage and with fewer error values. when this experiment is compared with the label and to the result achieved these algorithms predict well.

IV. PROPOSED WORK

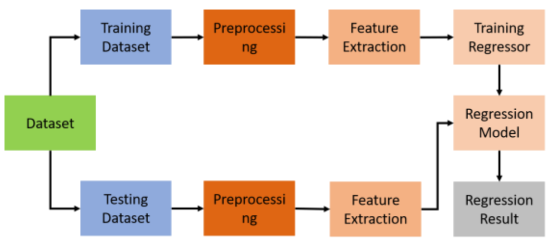
The purpose of this system is to determine the price of a house by looking at the various features which are given as input by the user. These features are given to the ML model and based on how these features affect the label it gives out a prediction. This will be done by first searching for an appropriate dataset that suits the needs of the developer as well as the user. Furthermore, after finalizing the dataset, the dataset will go through the process known as data cleaning where all the data which is not needed will be eliminated and the raw data will be turned into a .csv file. Moreover, the data will go through data preprocessing where missing data will be handled and if needed label encoding will be done. Moreover, this will go through data transformation where it will be converted into a NumPy array so that it can finally be sent for training the model. While training various machine learning algorithms will be used to train the model their error rate will be extracted

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

and consequently an algorithm and model will be finalized which can yield accurate predictions.

Users and companies will be able to log in and then fill a form about various attributes about their property that they want to predict the price of. Additionally, after a thorough selection of attributes, the form will be submitted. This data entered by the user will then go to the model and within seconds the user will be able to view the predicted price of the property that they put in.

4.1 Block Diagram of the System



The above block diagram is the traditional Machine Learning Approach. It consists of two sections: the training and the testing. The training has the following components: the label, input, feature extractor, and the machine learning algorithm. The testing section has the following components in it: the input, feature extractor, the regression model, and the output label.

Input: The input consists of data collected from various sources.

Feature Extractor: Only important features which affect the prediction results are kept. Other unnecessary attributes are discarded, like ID or name.

Features: After feature extraction only, some inputs are considered which largely contribute to the prediction of the model.

Machine Learning Algorithm: The ML Algorithm is the method by which an AI system performs its task, and is most commonly used to predict output values from given input values. Regression is one of the main processes of machine learning.

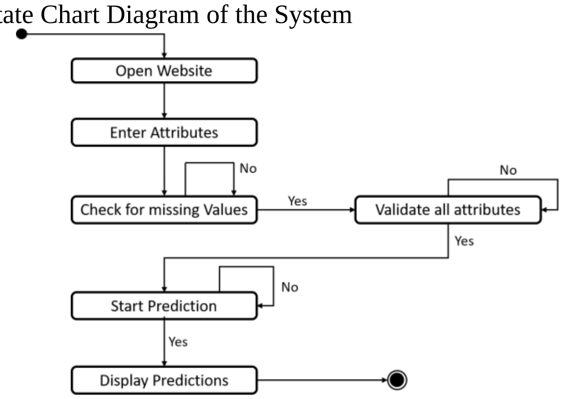
The Regression Model: The regression model consists of a set of machine-learning methods that allow us to predict a label variable (y) based on the values of one or more attribute/feature variables (x).

Briefly, the goal of a regression model is to build a mathematical equation that defines y as a function of the x variables.

Label: The label is the output obtained from the model after training.

The data obtained from the dataset is given as a training input first and the relevant training features are extracted. These training features are preprocessed to get a normalized dataset and labeling of the data row is done. The result from the training dataset is fed to the machine learning algorithm. The result from the Machine Learning Algorithm is fed to the Regression model, thus producing a trained model or trained regressor. This trained regressor can take the new data that is the extracted feature from the test as input and predict its output label.

4.1 State Chart Diagram of the System



The user will open the website and enter all the features/attributes of the house they wish to predict the price for. Furthermore, after the user clicks submit attributes will be checked for null values then all the attributes will be validated to check if they are in the same data type as necessary. Finally, after all the conditions are satisfied the data will be sent for prediction and the predicted price will be displayed to the user on the website.

V. IMPLEMENTATION

5.1 Datasets

We have used two datasets in this paper where various existing machine learning algorithms are applied to the datasets for predicting prices.

A. The first dataset is from the UCI Machine Learning Repository which concerns housing values in the suburbs of Boston. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. As this paper uses machine learning for price prediction, attribute variables are used to predict the label/price. The following table shows the set of attribute variables to develop the prediction model. This study uses 13 attributes as independent variables for predicting house prices.

Table 1: Attributes and label in the dataset (Boston)

Attributes	Description
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63) ^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $10,000's

B. This data was scraped from publicly available results posted every week from Domain.com.au by Anthony Pino. We have used 13 major attributes affecting the price for prediction.

Table 2: Attributes and label in the dataset(Melbourne)

Attributes	Description
CouncilArea	Governing council for the area
Method	Method of sale
Regionname	General region
Rooms	Number of rooms:
Type	Type of house
Distance	Distance from CBD in Kilometres
Bedroom2	Scraped # of Bedrooms (from different source)
Bathroom	Number of Bathrooms
Car	Number of carspots
Landsize	Land Size in Metres
BuildingArea	Building Size in Metres
YearBuilt	Year the house was built
Propertycount	Number of properties that exist in t suburb.
Price	Price in Australian dollars

5.1 Data Cleaning

Handling Missing Values

The Boston dataset only had five missing values. However, the Melbourne dataset had a lot of missing values (in thousands). Dropping the null values was not an option since it negatively affected the accuracy. These null values were handled by replacing them with the median value of the column. The replacement was done by implementing the simple imputer function in the pipeline itself. So that any missing value in the future would be handled as soon as the data passes through the pipeline. Additionally, the Melbourne dataset had missing label/price values which had to be dropped for better results.

Creating a pipeline and dropping columns

```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler # FOR SCALING
my_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy="median")),
    # We can add as many as we want in our pipeline
    ('std_scaler', StandardScaler()),
])
```

Figure 5.1.c Simple Imputer Code Snippet

We have dropped a few columns like ID or seller_info since they didn't have any effect on the final predictions.

Dropping Null Price Rows

```
[27] strat_test_set = strat_test_set.dropna(subset=["Price"]) #Option 1
     strat_test_set.shape

     (4929, 14)

[28] strat_train_set = strat_train_set.dropna(subset=["Price"]) #Option 1
     strat_train_set.shape

     (19660, 14)
```

Figure 5.1.d Dropping Null Price rows from Melbourne dataset

5.2 Data Preprocessing

Train and Test Split

We have split the dataset into two sets i.e. the Training set and the Testing set. Training set consists of 80% of the dataset and the testing set has 20% of the dataset. We had columns with only two distinct values and wanted to make sure that the splitting should split these values in equal proportions. Therefore, we used a stratified shuffle split for train test splitting for better results.

```
from sklearn.model_selection import train_test_split
train_set, test_set = train_test_split(housing, test_size=0.2, random_state=42)
print(f"Rows in train set: {len(train_set)}\nRows in test set: {len(test_set)}\n")

Rows in train set: 404
Rows in test set: 102
```

Stratified Shuffle Split of CHAS

```
from sklearn.model_selection import StratifiedShuffleSplit
split = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
for train_index, test_index in split.split(housing, housing['CHAS']):
    strat_train_set = housing.loc[train_index]
    strat_test_set = housing.loc[test_index]

strat_test_set["CHAS"].value_counts()

0    95
1     7
Name: CHAS, dtype: int64

strat_train_set["CHAS"].value_counts()

0    376
1     28
Name: CHAS, dtype: int64
```

Figure 5.2.a Stratified Shuffle Split in Boston dataset

```
[25] from sklearn.model_selection import train_test_split
     train_set, test_set = train_test_split(housing, test_size=0.2, random_state=42)
     print(f"Rows in train set: {len(train_set)}\nRows in test set: {len(test_set)}\n")

     Rows in train set: 25136
     Rows in test set: 6284
```

Stratified Shuffle Split of "Type"

```
[26] from sklearn.model_selection import StratifiedShuffleSplit
     split = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
     for train_index, test_index in split.split(housing, housing['Type']):
         strat_train_set = housing.loc[train_index]
         strat_test_set = housing.loc[test_index]
```

Figure 5.2.b Stratified Shuffle Split in Melbourne dataset

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

5.3 Correlation and Data Visualization

Correlation between all attributes and the label:

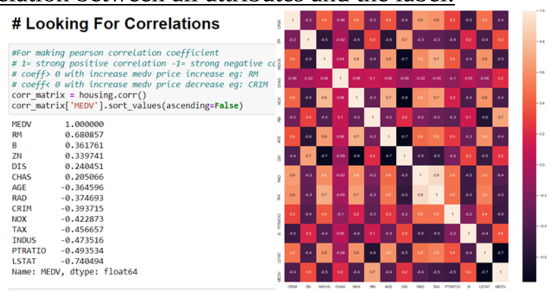


Figure 5.3.a Correlation heat map between all attributes and the label (Boston)

From the above figure, we can see that RM (No of rooms) was highly positively correlated followed by B and LSTAT and PTRATIO were the two most negatively correlated attributes. This means that if the value of RM or B has increased the price will increase and if the PTRATIO of LSTAT were to increase the price would decrease.

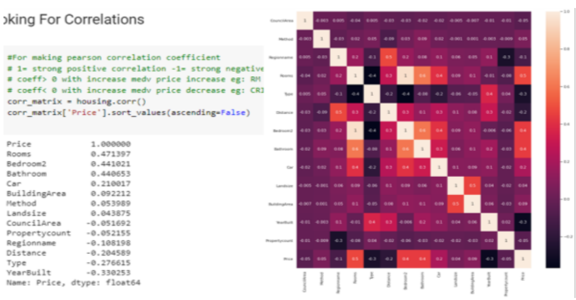


Figure 5.3.b Correlation heat map between all attributes and the label (Melbourne)

From the above figure, we can see that the attribute Rooms were highly positively correlated followed by Bedrooms and Type and YearBuilt were the two most negatively correlated attributes. This means that if the value of Rooms or Bedrooms were increased then the Price will increase and if the Type of YearBuilt were to increase the price would decrease.

Determining the relationship between Label and one random Attribute:

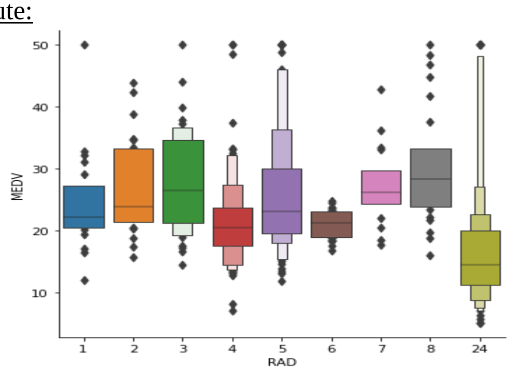


Figure 5.3.c MEDV(price) wrt RAD

For a maximum of the Distance's the inner 50% of the price was between 15 to 35 and the median hitting mostly at 20. However, for RAD 21 the price experienced a downfall as the distance from the highway increased the price got lower. Median staying at just 15.

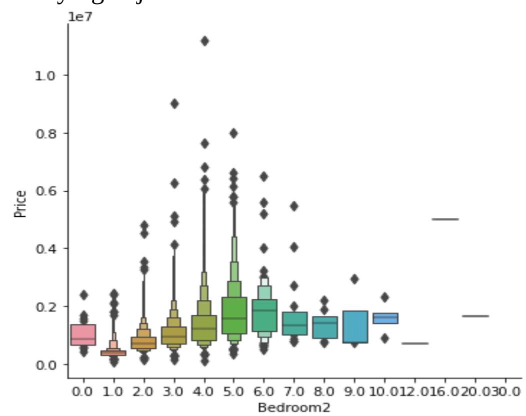


Figure 5.3.d Price wrt Bedroom count

It was observed that increasing bedroom count did not affect the price that much and surprisingly houses with several bedrooms between 3 to 6 topped the chart rather than the houses which had 8 to 30 bedrooms. And a house with 4 bedrooms had the highest price as compared to others.

VI. RESULTS

The following two tables show the RMSE Scores (mean and standard deviation) and Mean Cross-Validation Scores of the Boston dataset for four different machine learning algorithms and the Melbourne dataset for five different machine learning algorithms. Using these tables, we can infer the accuracy of all the algorithms for both the datasets and determine the best suitable algorithm for price prediction.

Boston Dataset			
Algorithm	**RMSE mean**	**RMSE Std. Dev.**	**Mean CrossVal Score**
XGBoost	3.06	0.75	0.88
Random Forest	3.38	0.76	0.85
Decision Tree	4.189	0.848	0.76
Linear Regression	4.22	0.75	0.69

Table 6.1 RMSE and Mean CrossVal of Boston Dataset

In general, the best accuracy was provided by the XGBoost Regression machine learning algorithm closely followed by Random Forest regression algorithm and Decision tree coming at the third place with a sizable difference. Additionally, the Linear Regression algorithm coming at last with a huge gap if compared to XGBoost. All these model predictions were checked for overfitting by evaluating these values using the cross-validation technique. So, these scores are extremely accurate.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

Melbourne Dataset			
Algorithm	**RMSE mean**	**RMSE Std. Dev.**	**Mean CrossVal Score**
XGBoost	299880.4	23019.5	0.79
Random Forest	312347.6	19880.3	0.76
Gradient Boosting	342911.4	26544.5	0.73
Decision Tree	421026.9	15907.3	0.6
Linear Regression	519029.1	97954.6	0.38

Table 6.2 RMSE and Mean CrossVal of Melbourne Dataset

In general, the best accuracy was provided by the XGBoost Regression machine learning algorithm closely followed by Random Forest regression algorithm and Gradient Boosting algorithm coming at the third place with a slight difference. Additionally, the decision tree algorithm came fourth with a sizable difference in scores, and finally, the Linear Regression algorithm came last with a huge gap if compared to XGBoost. All these model predictions were checked for overfitting by evaluating these values using the cross-validation technique. So, these scores are extremely accurate.

6.1 Price Predicting Website

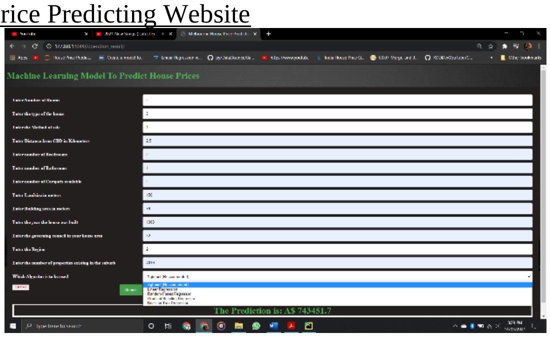


Figure 6.1

This is our take at developing a real-life solution for a house price prediction query. By developing a clean and simple UI, individuals with little to no experience will easily be able to predict prices for their dream house. This website was built on Flask technology and has a form where users can enter all the features of their house and predict prices as per their convenience. Additionally, the website provides users with the option of using the various machine learning algorithms for prediction (the best one comes auto-selected).

VI. CONCLUSION AND FUTURE SCOPE

Buying your own house is what every human wish for. Using this proposed model, we want people to buy houses and real estate at their rightful prices and want to ensure that they don't get tricked by sketchy agents who just are after their money. Additionally, this model will also help Big companies by giving accurate predictions for them to set the pricing and save them from a lot of hassle and save a lot of precious time and money. Correct real estate prices are the essence of the market and we want to ensure that by using this model.

The system is apt enough in training itself and in predicting the prices from the raw data provided to it. After going through several research papers and numerous blogs and articles, a set of algorithms were selected which were suitable in applying on both the datasets of the model. After multiple testing and training sessions, it was determined that the XGBoost Algorithm showed the best result amongst the rest of the algorithms. The system was potent enough for Predicting the prices of different houses with various features and was able to handle large sums of data. The system is quite user-friendly and time-saving.

The supplementary feature that can be added to our proposed system is to avail users of a full-fledged user interface so there can be multiple functionalities for users to use with the ML model for numerous locations. Also, an Amazon EC2 connection will take the system even further and increase the ease of use. Lastly, developing a well-integrated web application that can predict prices whenever users want it to will complete the project.

VIII. REFERENCES

[1] Thuraiya Mohd, Suraya Masrom, Noraini Johari, "Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia ", International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, 2019.

[2] G. 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.

[3] 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

[4] 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.

[5] Jingyi Mu, Fang Wu,and Aihua Zhang , " Housing Value Forecasting Based on Machine Learning Methods", Hindawi Publishing Corporation Abstract and Applied Analysis, Volume 2014.

[6] Sayan Putatunda, "PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market".

[7] Atharva Chouthai, Mohammed Athar Rangila , Sanved Amate, Prayag Adhikari, Vijay Kukre , "House Price Prediction Using Machine Learning" , International Research Journal of Engineering and Technology(IRJET), Vol:06 Issue: 03, 2019.

[8] B.Balakumar, P.Raviraj, S.Essakkiammal , "Predicting Housing Prices using Machine Learning Techniques".

[9] Akshay Babu, Dr. Anjana S Chandran , "Literature Review on Real Estate Value Prediction Using Machine Learning" , International Journal of Computer Science and Mobile Applications, Vol: 7 Issue: 3, 2019.

[10] Mr. Rushikesh Naikare, Mr. Girish Gahandule, Mr. Akash Dumbre, Mr. Kaushal Agrawal, Prof. Chaitanya Manka , "House Planning and Price Prediction System using Machine Learning" , International Engineering Research Journal, Vol:3 Issue: 3, 2019.

[11] Aswin Sivam Ravikumar, Thibaut Lust, "Real Estate Price Prediction Using Machine Learning", 2016.

[12] Bindu Sivasankar, Arun P. Ashok, Gouri Madhu, Fousiya S , "House Price Prediction" , International Journal of Computer Science and Engineering(IJCSE), Vol: 8 Issue: 7, 2020.

[13] M Thamarai, S P Malarvizhi, " House Price Prediction Modeling Using Machine Learning", International Journal of Information Engineering and Electronic Business(DJIEEB), VoL12, No.2, pp. 15-20, 2020. DOI: 10.5815/ijieeb.2020.02.03