Analysing the Features to Predict Housing Prices

Chandana Haluvarthi Prabhudeva
MSc in Data Analytics
National College of Ireland
Dublin, Ireland
X22167099@student.ncirl.ie

Pratheek Gogate
dept. Data Analytics
National College of Ireland
Dublin, Ireland
X22159789@student.ncirl.ie

***Abstract*—Housing market is the most dynamic industry where prediction of the prices accurately is the most crucial part for various of the stakeholders like buyers, sellers, real estate agents. The main aim of this project is to predict the housing prices based on the details of the house in a dataset which consists of the area per sqft, number of bedrooms, location and other important features .We are applying machine learning algorithms to to better predict the prices so that the prediction of prices will help the stakeholders to not overpay for the property or to get undervalue a property while selling it. The Linear regression is applied on the housing dataset and have got 76 percent R-Square after applying the optimisation techniques on the code.**

***Index Terms*—House Price, Linear Regression, Stakeholders**

I. Introduction

Real estate market has always been an critical factor of the economy, it reflects many facts such as potential investments, financial security it could also be a desire to have a home for themselves. Owing a house is a major mile stone for this generation of people as it is influenced my many factors, it is sometimes considered a measure to gauge your success.

In this realm of real estate, where buying and selling house is a major financial decision for individuals such as buyers and sellers, prediction of the house price is of very crucial importance.The goal of this project is to use the machine learning to increase the precision of prediction of home prices, which will ultimately lead to better decision-making for stakeholders in the housing business.

The accuracy of price projections may have a significant influence on both financial results and strategic decisions, even for a first-time homeowner looking for affordability and a real estate expert figuring out appropriate listing pricing, or an investor looking for possibilities to invest in real estate. The physical characteristics of a property, such as its size and condition, as well as its location, accessibility to amenities, and current market trends, all have an impact on how much a house costs.

By using machine learning algorithms to examine historical data and find patterns that may not be seen using conventional methods, this study looks forward to close this gap. The research intends to develop prediction models that provide more precise estimations by training algorithms to recognise correlations between different parameters and the final sale price and help the stakeholders make a better and wise decisions while they do their business in real estate.

II. Literature Review

Buying a house is one of the biggest event in any person's life because it is not something that happens regularly and usually that happens only once in their life. So it is very important to do proper budgeting and plan the requirements of a house and then invest in it, our study is going to help such people.

The Audience for Paper [1] is non-householders as it is going to help them to find the best house based on their requirement and their budget. During the research, they used different regression techniques, including Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting, and Ada Boost Regression, for house price predictions. They used a dataset that had merchandise, fare ranges, and market trends to estimate calculated prices. This research helped two types of people, first one is buyers as it will help them to predict the optimal time to purchase a house, and the other one is the seller for whom it helps them to accurately fix the best possible rate based on different factors such as physical conditions, concept, and location.

Actually, there is a rumor about real estate that usually pricing is done on hype rather than estimating the price based on the infrastructure it has. It is true to an extent because there is no transparency. The objective of paper [2] is to bring that transparency in predicting house prices based on real factors. The evaluation was done- based on fundamental parameters that are mostly considered when determining housing prices. They used Multiple regression models to achieve an accurate result and the final prediction was performed using a weighted mean of these models and the result indicated that this approach had a minimal error rate compared to individual techniques.

In this 21st century, it is very difficult to buy a house in any part of the world because of the high cost and more importantly there is no proper guidance on how much can be spent to buy a house based on the facilities it has. Paper [3] is specifically about a small town in the West Godavari district of Andhra Pradesh. The objective of this research was to predict the price of a house due to the demand it has. While doing the research they used factors like the number of bedrooms, house age, proximity to transportation, schools, and shopping malls, to model house availability to predict the house price which is fair enough considering these are the things usually which are part of Buyer's wish list and they are the once who are ultimately taking the advantage of these things. To predict the house price they used decision tree classification, decision tree regression, and multiple linear regression techniques and also they used the Scikit-Learn Machine Learning tool to build the models and analyse the results.

Machine Learning has played a major role in the fields like image detection, spam recognition, speech commands, product recommendations, and medical diagnosis. Paper [4] focuses on predicting house prices using machine learning. In this research, they have used different techniques and compared and explored various prediction methods, and finally selected lasso regression for its adaptable and probabilistic methodology in model selection. This Model gave better results than already existing other models. In the research, they also used XGBoost, lasso regression, and neural networks for housing price prediction and while evaluating Lasso Regression, outperformed other models. This research would have helped house sellers and real estate agents to evaluate house prices in a better way.

III. METHODOLOGY

In any project, the secret behind completing a project successfully is proper planning and the steps it takes to complete it. Those steps need to be performed in proper order and then only the desired result can be obtained. For example, if we are cooking something and we have all the resources and we will get the proper result only if we follow the steps in proper order.

Similarly in a Data Mining project, this holds good. Usually in a Data-Driven project, we use Cross-Industry Standard Process for Data Mining (CRISP-DM). It provides a structured approach to guide data scientists and analysts in the different stages of a data mining project. It has six stages in it. Figure 1 shows the different stages of this approach.

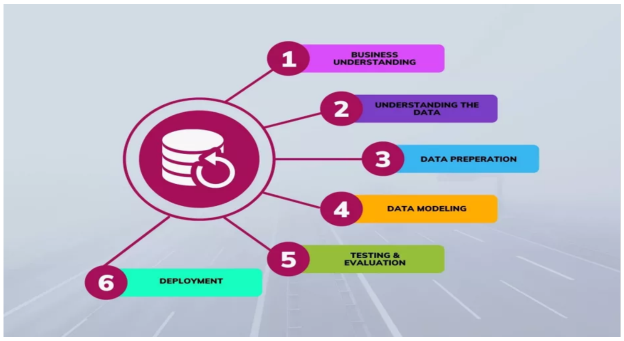


Fig. 1. CRISP_DM

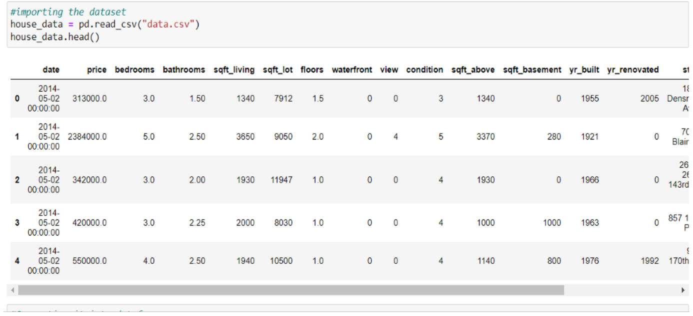
A. Business Understanding:

The First phase of a project is understanding what the target is and what resources we have. Usually before starting any project the goal will be clearly mentioned and what resources can be used and what exactly we need to solve will be defined before taking up the project the one who is going to be part of the project needs to understand these things clearly because if they don't understand these things project may not go as per the plans and ultimately it may fail.

In our project, There is a House Price prediction dataset and it has columns such as date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, sqft_above, sqft_basement, yr_built, y_renovated, street, city, state zip, country by using which we need to predict the house price which will ultimately help the buyers and sellers of a house. With this information, buyers can plan and decide at what time they can buy the house and at what location they can buy it, and at what price they will get a house if they have certain requirements. Sellers or builders can decide if is the right time to build the house, whether is there a demand for the house in the market currently, and at what price they can sell the house based on the features it has and many more.

B. Data Understanding

Once we understand what our goal is, the next step is data collection from the sources and understanding that data. Understanding the data includes what are features it has, which variable can be considered as the target variable, is the data is good enough to analyse, whether is there any missing or duplicate data, and also to check if there are any patterns in the data which can be helpful while building the models or something which is going to help the business.



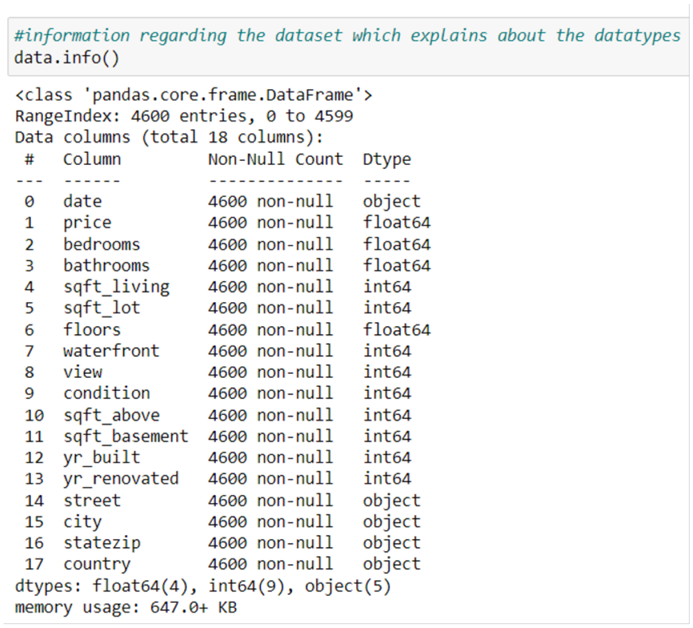
```
#importing the dataset
house_data = pd.read_csv("data.csv")
house_data.head()
```

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated	st
0	2014-05-02 00:00:00	313000.0	3.0	1.50	1340	7912	1.5	0	0	3	1340	0	1955	2005	18 Densr A
1	2014-05-02 00:00:00	2384000.0	5.0	2.50	3650	9050	2.0	0	4	5	3370	280	1921	0	70 Blair
2	2014-05-02 00:00:00	342000.0	3.0	2.00	1930	11947	1.0	0	0	4	1930	0	1966	0	26 26 143rd
3	2014-05-02 00:00:00	420000.0	3.0	2.25	2000	8030	1.0	0	0	4	1000	1000	1963	0	857 1 P
4	2014-05-02 00:00:00	550000.0	4.0	2.50	1940	10500	1.0	0	0	4	1140	800	1976	1992	9 170th

Fig. 2. Data Set table

In our project, we have a dataset where we need to predict the price of a house based on different features. So, we can consider price as our target variable and all the other input variables. In which Date indicates the date on which the data was recorded, Price indicates the actual price of the house, bedrooms, and bathrooms indicate the number of bedrooms and the number of bathrooms it has in a house respectively, Sqft_living indicates The total living area in square feet, Sqft_lot indicates the total area of the lot in square feet, which represents the land area of the property, Floors indicates the number of floors in a house, Waterfront indicates a binary value of 1 or 0 based on whether it has a view of water body, Condition represents the health of a house and it is measured using rating, and all other variables are self-explanatory. Figure 2 shows the dataset with the top 5 values as we have used the head function.

As a programmer, it is very essential to use correct data types for particular variables. Usually, the numerical values will be represented using either Integer or Float and then non-numerical values will be represented using objects. In our case, we should check all the variables carefully because we should transform accordingly before plotting them in a map or while model building. Figure 3 shows the datatypes of all the variables and shows the number of rows it has.



```
#information regarding the dataset which explains about the datatypes
data.info()
```

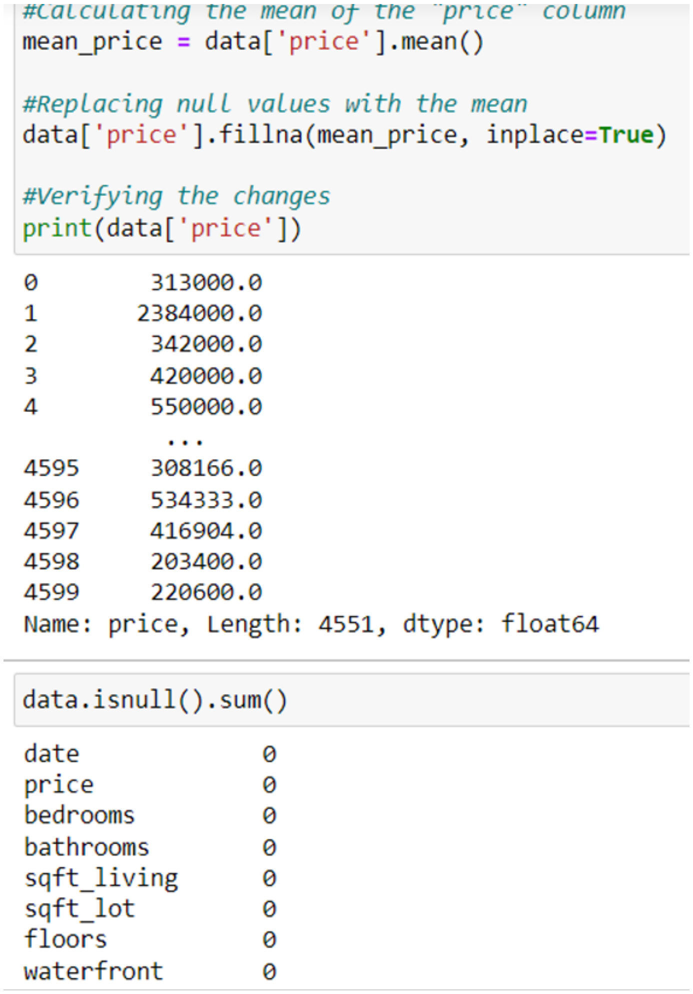
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4600 entries, 0 to 4599
Data columns (total 18 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   date           4600 non-null   object 
 1   price          4600 non-null   float64
 2   bedrooms       4600 non-null   float64
 3   bathrooms      4600 non-null   float64
 4   sqft_living    4600 non-null   int64  
 5   sqft_lot       4600 non-null   int64  
 6   floors         4600 non-null   float64
 7   waterfront     4600 non-null   int64  
 8   view           4600 non-null   int64  
 9   condition      4600 non-null   int64  
 10  sqft_above     4600 non-null   int64  
 11  sqft_basement  4600 non-null   int64  
 12  yr_built       4600 non-null   int64  
 13  yr_renovated   4600 non-null   int64  
 14  street         4600 non-null   object 
 15  city           4600 non-null   object 
 16  statezip       4600 non-null   object 
 17  country        4600 non-null   object 
dtypes: float64(4), int64(9), object(5)
memory usage: 647.0+ KB
```

Fig. 3. Data Set Description

C. Data Preparation

Data preparation has its own significance in any data-driven project as it is important to clean and make the necessary transformation before building the model. In our project, we initially checked if there are any duplicate values present and we did not have any duplicate values. Then we checked for null values, and we found that there were 49 null values for the price. Since Price was a numerical value and as most of the values were near to the mean value and hence, we replaced the null values with mean values which we could see in Figure 5. We have also transformed some of the variables.

Price, bedrooms, bathrooms, and floors have been converted to an integer data type as they were in float previously and street, city, state zip, and country have been changed to string as they were in object type also datatype of date was changed to date as it was in object type previously and it is not feasible to have a date in that format. Also, we checked for a correlation between input variables and there was some correlation but that was ignored after having sufficient knowledge in this particular domain.



```
#Calculating the mean of the "price" column
mean_price = data['price'].mean()

#Replacing null values with the mean
data['price'].fillna(mean_price, inplace=True)

#Verifying the changes
print(data['price'])
```

```
0        313000.0
1       2384000.0
2        342000.0
3        420000.0
4        550000.0
          ...
4595     308166.0
4596     534333.0
4597     416904.0
4598     203400.0
4599     220600.0
Name: price, Length: 4551, dtype: float64
```

```
data.isnull().sum()
```

```
date             0
price            0
bedrooms         0
bathrooms        0
sqft_living      0
sqft_lot         0
floors           0
waterfront       0
```

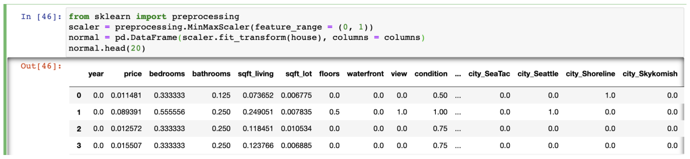
Fig. 4. Data Cleaning

D. Model Normalization

Model optimization is an important step in machine learning model development, as it fine tunes the various aspects and helps in achieving the best possible output for the given model by giving higher accuracy, RMSE value and also helps in tuning under fitting and over fitting of the parameters.

To optimize our linear regression model we have used normalization technique called minmax scaler, which transforms all the features in the dataset to fall in a specific range, most likely that range is between o and 1. This step is achieved by deleting the minimum value from the feature and dividing it by the range(max - min).

In our code we have applied minmaxscaler to the house dataframe, and the normal dataframe which is resulted has scaled between the range of 0 and 1, this has reduced any of the larger magnitude of the data from dominating the machine learning models and helps in ensuring that all the features that are used in our linear regression model are given the same importance and the model is not sensitive to any of the features.



```
In [46]: from sklearn import preprocessing
         scaler = preprocessing.MinMaxScaler(feature_range = (0, 1))
         normal = pd.DataFrame(scaler.fit_transform(house), columns = columns)
         normal.head(20)
Out[46]:
```

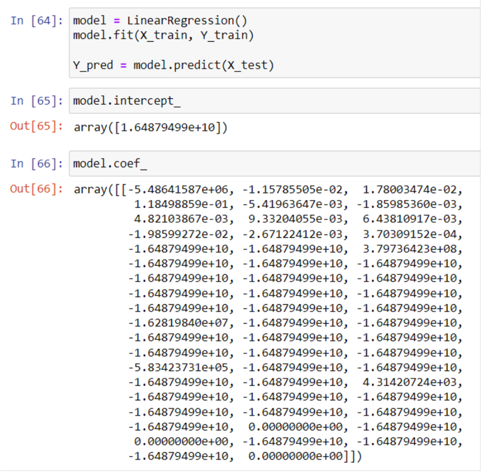
	year	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	...	city_SeaTac	city_Seattle	city_Shoreline	city_Skykomish
0	0.0	0.011481	0.333333	0.125	0.073652	0.006775	0.0	0.0	0.0	0.50	...	0.0	0.0	1.0	0.0
1	0.0	0.089391	0.555556	0.250	0.249051	0.007835	0.5	0.0	1.0	1.00	...	0.0	1.0	0.0	0.0
2	0.0	0.012572	0.333333	0.250	0.118451	0.010534	0.0	0.0	0.0	0.75	...	0.0	0.0	0.0	0.0
3	0.0	0.015507	0.333333	0.250	0.123766	0.006885	0.0	0.0	0.0	0.75	...	0.0	0.0	0.0	0.0

Fig. 5. MinMAx scaler

E. Model Building

This is the most important step, and all the previous steps are the supporting steps for this. Based on the model built the result will help the Landlords, Builders, and Customers it will help them to take decisions about housing. Looking at the data we had the option to build the model with Linear Regression, XGBoost, Decision tree, Random Forest, and Support Vector Machine. But after doing a thorough analysis and Literature Review [1],[2],[3],[7], we proceeded with Linear Regression. So when we build the model we will get a predicted value of a house based on different features.

By using this customers will get to know about the pricing of the house that they are looking for and since this calculation is done using proper data and algorithms, there is no chance of cheating the people in house pricing as they will already know about the approximate price with some differences because of external issues and depending on the house owners. In below figure 6 we could see a model built using Linear Regression.

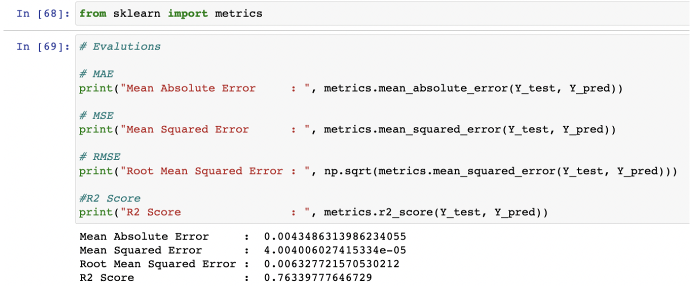


```
In [64]: model = LinearRegression()
         model.fit(X_train, Y_train)

         Y_pred = model.predict(X_test)

In [65]: model.intercept_

Out[65]: array([1.64879499e+10])

In [66]: model.coef_

Out[66]: array([[-5.48641587e+06, -1.15785505e-02,  1.78003474e-02,
                  1.18498859e-01, -5.41963647e-03, -1.85985360e-03,
                  4.82103867e-03,  9.33204055e-03,  6.43810917e-03,
                 -1.98599272e-02, -2.67122412e-03,  3.70309152e-04,
                 -1.64879499e+10, -1.64879499e+10,  3.79736423e+08,
                 -1.64879499e+10, -1.64879499e+10, -1.64879499e+10,
                 -1.64879499e+10, -1.64879499e+10, -1.64879499e+10,
                 -1.64879499e+10, -1.64879499e+10, -1.64879499e+10,
                 -1.64879499e+10, -1.64879499e+10, -1.64879499e+10,
                 -1.62819840e+07, -1.64879499e+10, -1.64879499e+10,
                 -1.64879499e+10, -1.64879499e+10, -1.64879499e+10,
                 -1.64879499e+10, -1.64879499e+10, -1.64879499e+10,
                 -5.83423731e+05, -1.64879499e+10, -1.64879499e+10,
                 -1.64879499e+10, -1.64879499e+10,  4.31420724e+03,
                 -1.64879499e+10, -1.64879499e+10, -1.64879499e+10,
                 -1.64879499e+10, -1.64879499e+10, -1.64879499e+10,
                 -1.64879499e+10,  0.00000000e+00, -1.64879499e+10,
                  0.00000000e+00, -1.64879499e+10, -1.64879499e+10,
                 -1.64879499e+10,  0.00000000e+00]])
```

Fig. 6. Linear Regression Model

IV. RESULTS AND EVALUATION

After doing the optimization by using the minmax scaler normalization technique, Linear regression model is built and the result is yield a precision of 76.3 percent.



```
In [68]: from sklearn import metrics

In [69]: # Evalutions

         # MAE
         print("Mean Absolute Error      : ", metrics.mean_absolute_error(Y_test, Y_pred))

         # MSE
         print("Mean Squared Error       : ", metrics.mean_squared_error(Y_test, Y_pred))

         # RMSE
         print("Root Mean Squared Error : ", np.sqrt(metrics.mean_squared_error(Y_test, Y_pred)))

         #R2 Score
         print("R2 Score                 : ", metrics.r2_score(Y_test, Y_pred))

         Mean Absolute Error     :  0.0043486313986234055
         Mean Squared Error      :  4.004006027415334e-05
         Root Mean Squared Error :  0.006327721570530212
         R2 Score                :  0.76339777646729
```

Fig. 7. Evaluvation Metrics

1. Mean absolute error(MAE):
The mean absolute error calculates the average absolute difference between the actual values and the expected values. The result is 0.00434 and it shows the average deviation between the prediction and the average result. As we can see that the value is low so we know that the difference is not huge, which indicates a that we have a good model.

2. Mean Squared Error (MSE):
The average of the squared discrepancies between the anticipated values and the actual values is computed using the mean squared error. The result is 4.004, this shows us that it is more sensitive to out liners, it could have been a little better.

3. Root Mean Squared Error (RMSE):
The square root of the Mean Squared Error is known as the Root Mean Squared Error. The result is 0.00632and Its meaning is the same as that of the target variable's original units. The average size of the forecast mistakes is conveyed by the RMSE and we can see that the error percent is less

4. R-squared (R2) Score:
The proportion of the dependent variable's (target) variation that can be predicted from the model's independent variables (features) is expressed statistically by the R-squared (R2) Score. The scale goes from 0 to 1, with 1 denoting that the model accurately predicts the target and we can see that our results are 0.763 which is a good score as it shows that the R-Square is 76 parentage.

The presented result indicates, as seen by the low error metrics and a relatively high R2 score, that the model has performed well in terms of predicting the target variable.

Predicting the house value or the price of the house will make a substantial benefit to many of the stakeholders in the real estate market, as we have different stakeholder we can see how it effects each of them with our accurate predictions:

1. Homebuyers:

Our results helps them to understand the clear picture of the market value of the house that they are planning to buy. By this they can avoid making an over payment for a property and make an appropriate negotiations with the sellers, by having an insite of the value of the house they can be more confident while doing their deals which will be profitable and they won't pifall for overpaying.

2. Sellers:

Our results help the sellers to list their house in the market for a competitive price as over price for a property can lead to not being bought by the holders as setting the correct by a data driven prediction will bring more buyers to the table and indeed help in selling the property sooner.

3. Real Estate Professionals:

Real estate agents and the brokers can use our prediction to provide their client with a well informed advice to buy the house, this help them helping their customers set a realistic prices in the market and avoiding of teh property being stale in market, and avoid the main problem of being high price and low in demand.

4. Government Agencies and Policy Makers:

The prediction of cost help the government agencies to make a better housing policies, that are inline with the current market trends. This will reduce the major issues like affordability issue and housing bubbles and makes sure that the market is stable.

V. CONCLUSIONS AND FUTURE WORK

The main reason for house price prediction is to make a decision while doing a business related to houses and make that decision a financially benefiting now, as it has major issue of pit-falling into losses.

Mainly a accurate prediction of the house leads to a empowerment of the stakeholders to have data driven in-site of the true value of the property, as this will be a huge financial savings but also a leads to a very transparent real estate business. With this predictions they can be confident on the decisions without any second guesses and align with their financial goals and needs.

We have got the R-Square of 76.3 percent for predicting the house prices, this can be improved with larger dataset and by also considering more factors such as the current market value of the land and the properties used for the construction of the house as these factors give more insite in the price prediction as they aid in deciding the estimated cost of the property.

REFERENCES

[1] House Price Prediction Using Regression Techniques: A Comparative Study CH. Raga Madhuri; G. Anuradha; M. Vani Pujitha, 2019 https://ieeexplore.ieee.org/abstract/document/8882834

[2] House Price Prediction Using Machine Learning and Neural Networks Ayush Varma; Abhijit Sarma; Sagar Doshi; Rohini Nair, 2018 https://ieeexplore.ieee.org/abstract/document/8473231

[3] House Price Prediction Modeling Using Machine Learning Dr. M. Thamarai, Dr. S P. Malarvizhi ,25 October 2019 http://www.mecs-press.net/ijieeb/ijieeb-v12-n2/IJIEEB-V12-N2-3.pdf

[4] House Price Prediction Using Machine Learning G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu, 2019 https://tinyurl.com/36sx4d86

[5] https://mymoodle.ncirl.ie/course/view.php?id=2093