Property Price Prediction Using Regression Algorithms, Chatbot Using NLP

Mayur Patel, Brandon Vinod, Vaibhavi Vaidya, Prof.Poonam Narkhede
Department of Computer, Mumbai University
Shivajirao S. Jondhale College of Engineering, Dombivli, India

Abstract- ***The role of computer has become more intrusive in the recent days. For our day to day task we use various technologies without even recognizing it. The field of machine learning, deep learning, AI aims to automate the system and provide fluid services to us with least human intervention. In this present paper we aim to provide prices of a property based on its features. Our aim is to deploy the model on internet so anybody can get information regarding the property they want. We used various machine learning regression algorithms to solve our use cases. In our case Random forest regression performed the best and gave us 92% accuracy. In this paper we present the working principle & basic concepts as well as application in various sectors in our case with Property Price Prediction. In our system we provide a efficient and accurate answer based on dataset of FAQ using NLP in AI. We used HTML, CSS, bootstrap, Flask for our front end application.***

Keywords- ***Linear Regression, Ridge regression, SVR,K-neighbour Regression, Random forest Regression, Ada boost regression, Gradient boost regression, XG Boost regressionn, machine learning, Chatterbot, corpus, NLP, Flask.***

I. INTRODUCTION

The process of learning or getting experience is nothing but going through past activities. By considering events that has happened in the past we can make a more conscious decision that may lead to correct decision making. This kind of behaviour is not only found in humans but also in animals. A chimp can display a wide range of emotions, and recognize them in mirror and can learn sign language. This all is possible because of learning. Similarly in case of machine learning we use past data to make a firm prediction regarding the future. There are two types of learning techniques in machine learning. They are Supervised machine learning and Unsupervised machine learning. Supervised machine involves learning from data which is well labelled and unsupervised machine learning involves learning from data which is not labeled.In unsupervised machine learning the main aim is to find pattern in the unlabeled data. In our case we are performing supervised machine learning. Price forecasting can also be done in the similar way by using the existing data. From the past data we can see the trends in the data and with this we can make a firm decision. Nowadays peoples interaction with machines have become common. Be it real estate agent or customer, real estate chatbot prove to be effective to both when it comes to saving time, money & additional resources. With a large number of property under preview, real estate receives a lot of enquiries so chatbot can instantly address such queries. Our aim is to create a front end app with machine learning regression model for price forecasting and chatbot using nlp for user assistance.

II. LITERATURE SURVEY

The value of real estate property is mostly affected by its location and area. However by playing with different attributes and try out some of the of them in our use cases proved to be useful. However they cannot be seen directly. We need to perform various EDA techniques with visualization using various data visualization techniques like bar plot, scatter plot etc. However EDA also involves various non graphical techniques for dealing with various factors in data, like dealing with outliers using z-score, IQR etc. This is explained in detail in the paper by Mathieu [1] with each and every use cases distinctly visualized. The importance of EDA and its effect on performance on machine learning algorithm is evident.EDA helps us to utilize the maximum information of the available data. However the main aim is to create a robust generalized model and for this we also need to know various techniques. The paper by Alisha [2] explains various terms like hyper parameter optimization and tweaking the machine learning algorithms to squeeze the maximum score. Their approach on the problem statement was inspiring and effective. The flow of development and the way to analyse the performance using various metrics was also explained.

Just like any emerging technology, chatbot will only become widely adopted if it's shown that they can solve real problems. Now that we understand the main problems consumers have with traditional online experiences, let's look at if (and how) chatbot can actually solve these problems. In our survey, we provided a brief description of how chat bots work and the types of tasks which is explained in paper by Mohammad Waseem[8]Every time a chatbot gets the input from the user, it saves the input and the response which helps the chatbot with no initial knowledge to evolve using the collected responses. With increased responses, the accuracy of the chatbot also increases.

III. SYSTEM DESIGN AND ARCHITECTURE

A. Data Collection

The first step in every use case for machine learning is to get the appropriate data set. This involves web scraping to collect data from different source from the internet. It can also be done using API, some sites provide API keys for data extraction. There are many open source public data libraries on internet like UCI machine learning repository, Google data set search engine , and Kaggle from where we found our required data.

B. Data Pre-processing

The data we get from various sources are not always ordered or structured, it may have missing values , it can have character data type. So before feeding it to the machine learning algorithm we need to clean the data and this is done using Exploratory data analysis techniques.

C. Training the models

After the above step is completed we divide our data into three sets mainly Training set, Test set, Validation set (Optional).We apply various machine learning models on training set and select the best models as our primary model. We then tune its parameter using various hyper parameter optimization techniques.

D. Testing and integration with UI

Once model is created it is tested on test set and validation set before deployment. We need to check for bugs before deploying the website on the server.

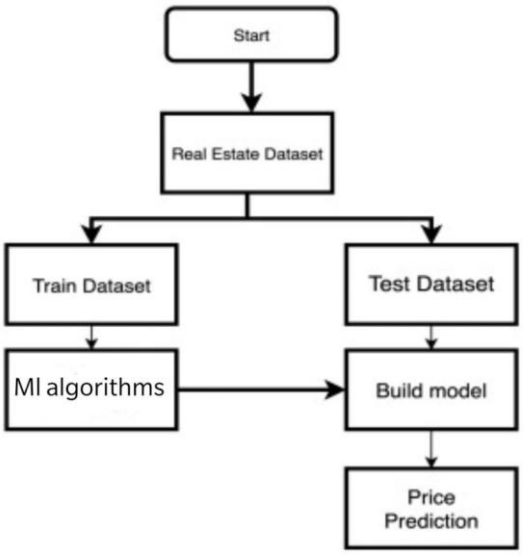


Fig 1. Generic Flow of Development

E. Natural Language Processing

Natural Language Processing, also known as NLP, is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully process large amounts of natural language data. Like the example with Amazon's Alexa assistant would be able to provide little to no value without Natural Language Processing (NLP).

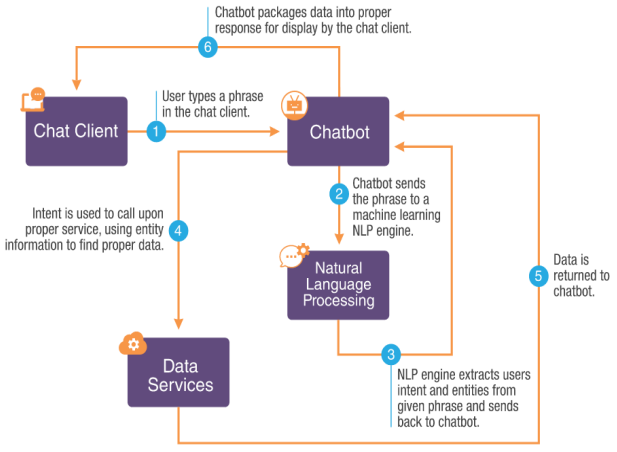


Fig 2. Chatbot Flow Diagram

IV. METHODOLOGY

A. Data Collection

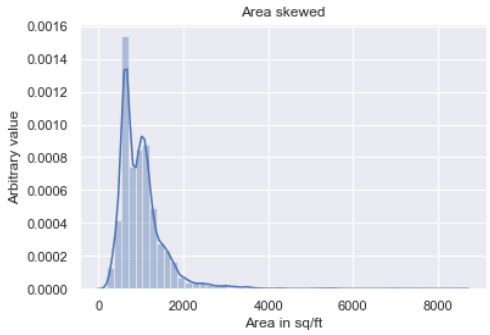
We wanted to build a model which can predict the prices of property of our locality. So for making such model we required data which has prices of property of our locality. We considered various data sets that were on the web. In the end we found a data which was perfect for our use case and had property prices of Mumbai, Thane and other locations. The data was uploaded on Kaggle[3]. The data had 6345 rows and 16 columns.

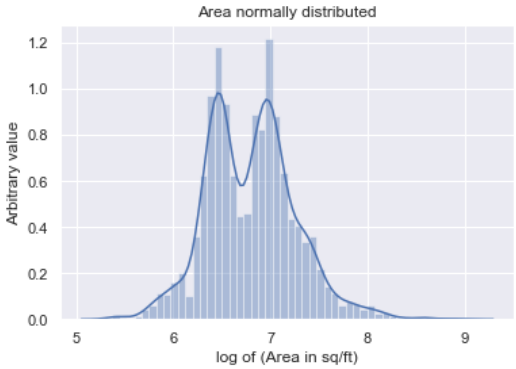
F. About Data

The data contained 6345 rows and 16 columns. The data had attributes like Area,Location,Number of rooms, Gymnasium availability, Lift availability, Car parking availability, Security availability, Children's garden availability, Clubhouse availability, Intercom connection availability, Garden availability, Indoor games availability, Gas pipeline connectivity, Jogging track availability, Swimming pool and the Price of the respective property. Out of the 16 attributes two are continuous attributes (Area, Price) and 13 attributes are discrete numerical attributes and one attribute is object type (Location).The data for the Chatbot is acquired from the dataset of property of the areas you are looking for in your budget.

G. *Data Pre-processing*

The attributes Price and Area were highly skewed. These attributes were continuous random variable. We converted them into Gaussian distribution or Normal distribution by log transformation .The distribution of data before and after log transformation is given below.





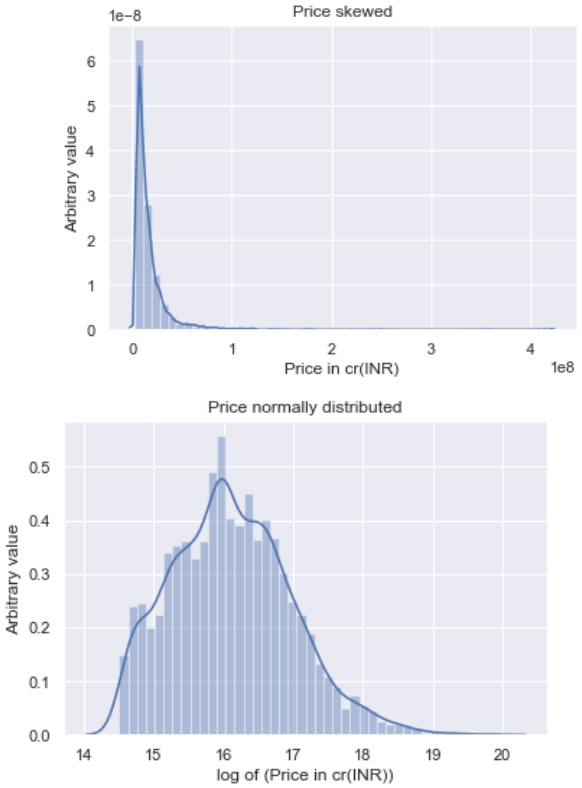


Fig3. Distribution of attributes Area and Price

This stage involves cleaning the data so that it can be given to a machine learning algorithm. First we check if there are any missing values in the data set and if present we use mean encoding if missing values are small and median encoding if outliers are present or we can also use random sample imputation or ken imputation. Machine learning algorithms only deal with numerical data. So we need to convert our categorical data, object data type into numerical data type. This can be done by using one hot encoding , target encoding ,mean encoding etc. In our case there were no missing values in our data set. However the location attribute had high cardinality. For dealing with high cardinal attribute we cannot use one hot encoding as one it leads to increase in numbers of columns and higher number columns leads to cruse of dimensionality which drops the accuracy of machine learning models. So in order to encode location in number we used a different approach where we first created a new attribute that was called per square foot and as the name suggests it has information about per square foot prices of the property. We obtained this attribute by dividing Price of property by Area of property. In order to encode location we formed a bucket of location based on their per square foot price and took the median of that bucket. One the basis of rank the location which has lowest per square foot price was labelled 0 and the highest was labelled 410 as there were 410 different locations in our data. This encoding method was effective as the lowest encoded location was evidently less priced in real life and the high encoded location was a very posh location in Mumbai. The effectiveness of the encoding can be proved with a visualization of scatter plot between Location and Price.

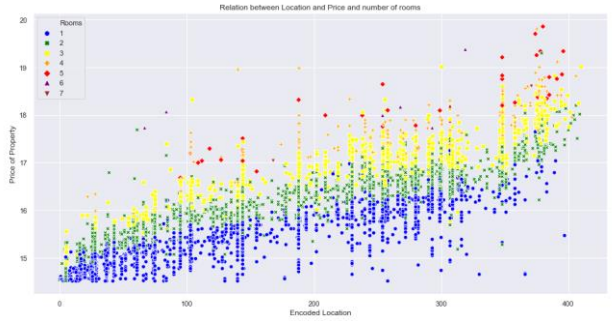


Fig4. Scatter plot Between Attributes Location and Price

The different symbols in the graph represents amount of rooms in the property.It is evident that price of property increase as we move from left to right on x-axis thereby proving that location effects the price of property.In the graph it is also scene that property having more number of rooms are much more costlier however after 4 to 5 count they are effectively same.

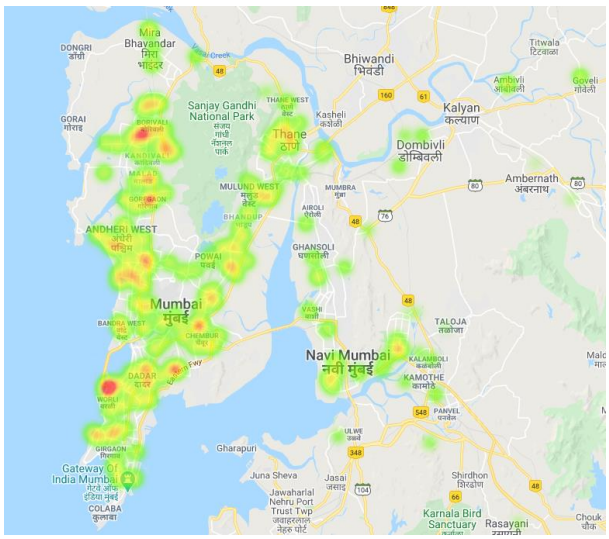


Fig5 . Heat map of price distribution based on location

The above figure is a heat map which helps us to visualize distributions of prices based on physical or actual locations.As you can see the locations in south Mumbai have red gradient specifying higher prices of property in that locality.While other properties have yellowish and green gradient which indicates lesser prices.From the map it is also clear that properties are build nearer to roads and railways as these factor affects the prices of property greatly!

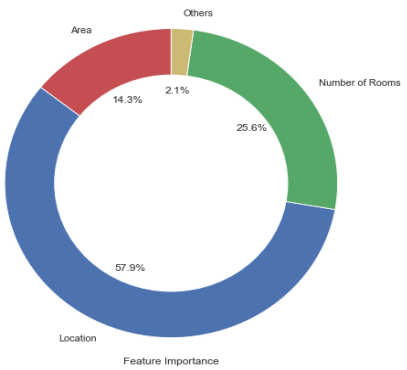


Fig 6. Feature importance

From the above pie chart we can say that the attributes Area , Location and Number of Rooms plays very vital role in predicting the price of the property[4]. While all other attributes only contribute 2.1 % as suggested by the pie chart .

H. Training the model

After data pre-processing we split our data into training set and test set with 70% training set and 30% test set. We now use various regression based machine learning algorithms like linear regression, Ridge regression, Random Forest regression, Ada boost regression, Gradient boost regression, XG boost regression. All the algorithms performed well however Random Forest and XG Boost came close .We performed hyper parameter tuning with randomized search in order to get best parameter for the machine learning algorithm. Chatbot is trained using a module known as Chatterbot corpus that helps to train the data.This is because each corpus is just a sample of various input statements and their responses for the boot to train itself with.

I. Results

We tried various mentioned machine learning algorithms [5] and the score (r2score) we got was is given below. These scores are mean of cross validation scores. The method we used for cross validation is KFold [6] as we are dealing with regression problem statement and it is much suitable for this use case.

Linear Regression: 0.910573103242809
Ridge Regression: 0.9105814781156265
SVR: 0.7426253391736471
K -Neighbour: 0.8601602072824667
Ada Boost: 0.8119448550989302
Gradient Boost: 0.9247354894991864
XG Boost: 0.9245608777424721
Random Forest: 0.9243831670710094

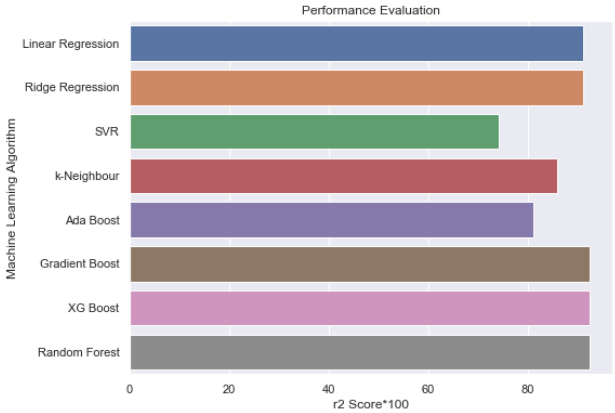


Fig7. Performance Evaluation

The scoring is done by using r2 score [7].This score is the average score given by the model using K fold cross validation techniques. We got low bias and low variance suggesting that model is generalized model.Random Forest regression , XG Boost regression , Gradient Boost regression performed equally well but we selected random forest as our final choice because we found its output for certain inputs were better compared to the other to algorithms.

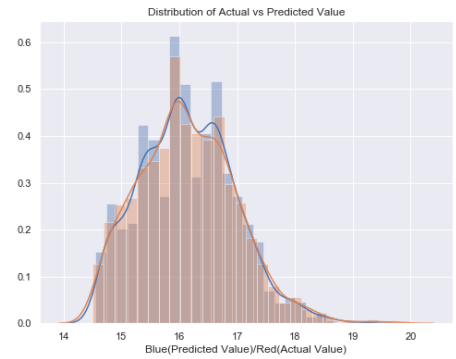


Fig 8. Distribution of actual vs. predicted values

From the graph above we can say that the prediction is actually good as it follows normal distribution or Gaussian distribution. There is almost not difference in distribution of data thereby confirming that model is performing well.

J. *Testing and integration with UI*

After creation of model and achieving good accuracy score, the next step is to deployment of model on a web page so anybody can use it. For deployment purpose we will use FLASK. It provides tools and technologies for developing a web application.We will make a Flask chatbot. Flask is a micro framework used for web development. We will follow the process given below:

1) Make a web app using the flask.
2) Make a directory for the templates.
3) Train the bot.
4) Make conversation with the bot.

So flask can carry out the entire all the back end task while we need to design front end using html, css and bootstrap.

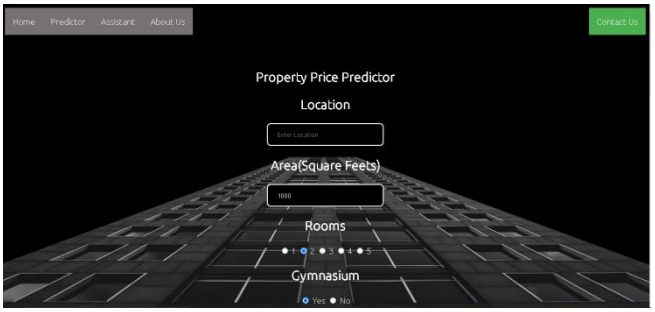


Fig 9. Output 1

Fig 10. Output 2

K. Future Scope

For the future we will consider to make our site much more user friendly, elegant and provide more information by suggesting users real estate properties based on their prediction. We would like to increase the area covers by our model and to make it nationwide. To improve data set by including much more information like nearest schools, hospital, shopping mall, air quality index and many more attributes. This way we can evaluate even more precisely as these factors affects the prices very effectively!. [10]Future intelligent chatbot should,

1) Implement improved NLP techniques

2) Learn to understand human context in conversation and respond accordingly with emotions or personalised content.

V. CONCLUSION

In this paper, we tried various regression based machine learning algorithms and came to conclusion that random forest regression performed best. Our data set contained various information regarding properties in Mumbai, Thane and other regions instead of simply the area, location and Price. The prices predicted by the model were extremely close to real prices listed online and know by the people. Most of the existing system does not have various attributes like we have in our data set making it much more reliable for the user. Every customer or user need appropriate answer and so database is used so that purpose can be solved. There are many NLP applications and programming interfaces and services that help in development of chatbot. Having the ability to improve itself with every interaction will likely improve the Chabot's capability of understanding the context of user's input, which would help the chatbot generate more accurate, relevant response. Our system gave and average accuracy of 92% which is considered great for predicting the prices of real estate properties.

VI. ACKNOWLEDGEMENT

We would like to thank the open source community for providing solutions to various problems we faced in our tasks. We sincerely wish to thank the project guide Prof.Poonam Narkhede for her encouraging and inspiring guidance helped us to make our project a success. Our project guide made sure we were on track at all the times with her expert guidance, kind advice and timely motivation which helped us to determine our project. We would like to thank our project coordinator Dr. Uttara Gogate for all the support she provided with respect to project. We also express our deepest thanks to our HOD Prof. Pramod Rodge whose benevolent helped us by making the computer facilities available and making it true success. Lastly, we would like to thank our college principle Dr. J.W.Bakal for providing lab facilities and permitting to go with our project. We would also like to thank our colleagues who helped us directly or indirectly during our project.

REFERENCES

[1] Alisha Kuvalekar "House Price Forcasting Using Machine Learning"

[2] Mathieu Komorowski "Exploratory Data Analysis"

[3] Data set https://www.kaggle.com/sameep98/housing-prices-in-mumbai

[4] Feature Importance: https://machinelearningmastery.com/calculate-feature-importance-with-python/

[5] Machine learning algorithms: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[6] KFold:https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

[7] R2score:https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

[8] CHATBOT LITERATURE SURVEY:https://www.edureka.co/blog/how-to-make-a-chatbot-in-python/

[9] CHATBOT OUTPUT PIC:https://laptrinhx.com/how-create-chatbot-in-few-minutes-using-python-or-flask-230723851/

[10] CHATBOT FUTURE SCOPE:https://www.callcentrehelper.com/ways-to-improve-chatbots-boost-satisfaction-124375.htm