House Price Prediction And Recommendation System

Abhishek Jadhav
dept. Information Technology (I2IT)
Pune, India
abhijadhav8308@gmail.com

Nikita Bharambe
dept. Information Technology (I2IT)
Pune, India
nikitavbharambe@gmail.com

Sharayu Saraf
dept. Information Technology (I2IT)
Pune, India
sharayusaraf28@gmail.com

Vedant Pathak
dept. Information Technology (I2IT)
Pune, India
vedant.pathak100@gmail.com

Prof. Vishal Chaudhary
dept. Information Technology (I2IT)
Pune, India
vishalc@isquareit.edu.in

International Institute of Information Technology, Pune

***Abstract*—We find that the positive effect of real estate investment on economic growth is stronger in short term while in long term turn to be negative. That's why, housing investment is an important factor for short-term economy fluctuations and lead to downside risk in the long term. This is an important motivating factor for house price prediction and recommendatin. Machine learning which plays an important role in various sector from past years. such as healthcare, finance, Manufacturing, Advertising, Real Estate, etc.**

In this project, we are going to create a web application which is used by two types of users. First is buyer, who will search house for buying purpose according to his requirements. And Second is seller, where seller have to add some property details for predicting the house price, enter date for forecasting the price till that date and budget range for recommending best location. This project uses datasets, which includes house's features, location and price. We are using different feature selection methods and feature extraction method with Multiple Linear Regression to predict the current house price and using content based recommendation system to recommend best location according to their budget in nearby area of interest.

I. Introduction

Prediction is an important part provided by Machine Learning. In machine learning model predictions allow businesses to make highly accurate guesses based on historical data. Prediction provide the business with insights that result in tangible business value. For example, if a model predicts a customer is probably going to churn, the business can target them with specific communications and outreach which will prevent the loss of that customer.

The aim of this system is to predict the house price and recommend the other houses according to their requirements. This is a website which will take users requirements as input and pass this input to the prediction module which used Catboost Algorithm for prediction. This system also allows user to forecast the predicted house price to particular date which is specified by user. This forecasting is done by using another time series model known as SARIMA (Seasonable Auto Regressive Integrated Moving Average Model) model. This system also includes recommendation system which suggest relevant items to the user. and the recommendation system system we are using Content Based filtering.

II. Literature Survey

A. Modelling House Price Prediction using multivariate analysis and Particle Swarm Optimization.

In this paper, some of the tests have been performed using linear regression and particle swarm optimization methods to perform house price prediction. The 1st approach is a quantitative prediction. A quantitative approach is an approach that is used to utilizes time-series data. The time series approach is to look for the relationship between current prices and prevalent prices. The 2nd approach is to use linear regression based on hedonic pricing. In linear regression, determine coefficients using the least square method, but it takes a long time to get the best formula. Particle swarm optimization is used to find the coefficients aimed at obtaining optimal results.

B. IEEE: House Price Prediction Using Various Regression Techniques: A Comparative Study

This paper mainly concentrates on the comparison between different machine learning algorithms. This paper involves predictions using different Regression techniques like Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting and Ada Boost Regression. House price prediction on a data set has been done by using all the techniques mentioned above, to find out the best among them. The aim of this paper is to help the seller to estimate the selling cost of a house perfectly and to help people to predict the exact time slap to accumulate a house. Some of the related factors which impacts the cost were also taken into considerations such as physical conditions, concept, requirements and location etc.

C. Real Estate Recommender System Using Case Based Reasoning Approach

This paper proposes a true Estate Recommender System using Case Based Reasoning Approach which may help customers to seek out a desired property. This proposed system uses a recommendation approach during look for property which assists the purchasers to seek out appropriate property and make decisions where they have the specified knowledge to judge a particular property. Furthermore, information available is extremely huge, therefore the recommender system assists the user to filter the available dataset consistent with user needs. Recommendation methods used for the searching is Case-Based filtering approach which can solve a new problem by retrieving the same problem that has been solved before and reuse the information which is used to solve this new problem. Also, system uses collaborative filtering approach which filters the properties supported other user rating for properties; the system will do recommendation supported the top-rated properties. Furthermore, system will recommend the user based on the most visited properties, where the system will count the number of visits to the database, after that based on the property with highest number of visits, system will recommend appropriate property to users.

D. IEEE: Predicting the Sales Prices of the Houses/ real estate property Using Regression Methods of Machine Learning

This paper consist solution for house's sales price prediction problem in which several standard machine learning algorithms with our original ideas like residual regressor, logit transform and neural network machine are combined. This find the solution for "House Prices: Advanced Regression Techniques" machine learning competition, which was held on Kaggle platform. The goal was to predict house's sale price by their features like house area, year of building, etc.

III. Design and Implementation

A. User Interface

The user interface of House Price Prediction and Recommendation System is web application. It includes two types of users. One is Buyer who want invest in property. And another is Seller. Seller must need to Sign up where the proccess for Sign up is not compulsory for Buyer. Both the users need to enter the details of property they want after that system will give the accurate predicted value. User can also be able to forcast the predicted value. In this project user also able to see the location as we are providing by using google API's.

B. Methodology

1) CatBoost Algorithm: CatBoost may be a machine learning algorithm that uses gradient boosting on decision trees. it's available as an open source library. Catboost are often used for solving problems, like regression, classification, multi-class classification and ranking. Modes differ by the target function, that we try to attenuate during gradient descend. Moreover, Catboost have prebuild metrics that are used to measure the accuracy of the model.

- CatBoost's main ability is to integrate a variety of differ- ent data types, such as images, audio, or text features into one framework but CatBoost also offers an idiosyncratic way of handling categorical data, requiring a minimum of categorical feature transformation, opposed to the majority of other machine learning algorithms,that cannot handle non-numeric values.
- CatBoost is built upon the basic concepts of decision trees and gradient boosting. Boosting algorithms work by sequentially combining multiple weak models thus making a strong predictive model.
- Because gradient boosting fits the decision trees sequen- tially, the fitted trees will learn from the mistakes of former trees and hence reduce the errors. This process of adding a new function to existing ones is continued until the selected loss function is no longer minimized.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (
catboost	CatBoost Regressor	64974.0859	14078810685.8847	118123.8016	0.8951	0.1685	0.1245	8.07
lightgbm	Light Gradient Boosting Machine	68839.2918	16321013769.4938	127379.1577	0.8781	0.1762	0.1315	0.28
xgboost	Extreme Gradient Boosting	70127.3391	16860578201.6000	129446.4461	0.8745	0.1800	0.1326	2.05
rf	Random Forest Regressor	71621.2060	18242493548.8560	134803.6375	0.8639	0.1830	0.1349	9.69
et	Extra Trees Regressor	72557.1988	18820921942.9531	136824.2531	0.8599	0.1861	0.1367	4.75
gbr	Gradient Boosting Regressor	81974.5643	20416905072.8391	142742.6270	0.8478	0.2027	0.1576	1.77
lasso	Lasso Regression	119369.1047	36869402009.6000	191888.2906	0.7243	0.3138	0.2325	0.29
llar	Lasso Least Angle Regression	119325.4894	36869253672.5004	191887.9021	0.7243	0.3128	0.2324	0.02
br	Bayesian Ridge	119421.4676	36873620089.0746	191900.4338	0.7243	0.3146	0.2328	0.07
lar	Least Angle Regression	119522.4089	36977078657.9807	192165.5122	0.7234	0.3153	0.2328	0.06
dt	Decision Tree Regressor	103452.0416	38499621173.6384	196174.5096	0.7133	0.2591	0.1907	0.39

Fig. 1. Comparison between CatBoost other Algorithm.

2) Time Series Model (SARIMA): The Seasonal Autoregres- sive Integrated Method (SARIMA) models the next step in the sequence as a linear function of the differenced observations, errors, differenced seasonal observations, and seasonal errors at prior time steps.

- A time series is a simple series of data points ordered in time. In a time series model ,the time is often the independent variable and the goal is usually to make a forecasting for the future.
- Time series analysis is a technique which deals with time series data, or trend analysis. What is Time series data? It means that data is in a series of particular time periods or intervals. Time series take many aspects into consid- eration such as Stationarity, Seasonality, Autocorrelation. There are many ways to predict time series predictions. For ex. Simple Moving Average, Autoregressive Moving Average, Exponential Smoothing, ARIMA.
- The algorithm we have used in our system is SARIMA which is an extension of ARIMA model that considers seasonal component while predicting the future price. notation for the model involves specifying the order for the AR(p), I(d), and MA(q) models as parameters to an ARIMA function and AR(P), I(D), MA(Q) and m parameters at the seasonal level, e.g. SARIMA(p, d, q)(P, D, Q)m where "m" is the number of time steps in each

season (the seasonal period). A SARIMA model can be used to develop AR, MA, ARMA and ARIMA models. The method is suitable for univariate time series with trend and/or seasonal components.

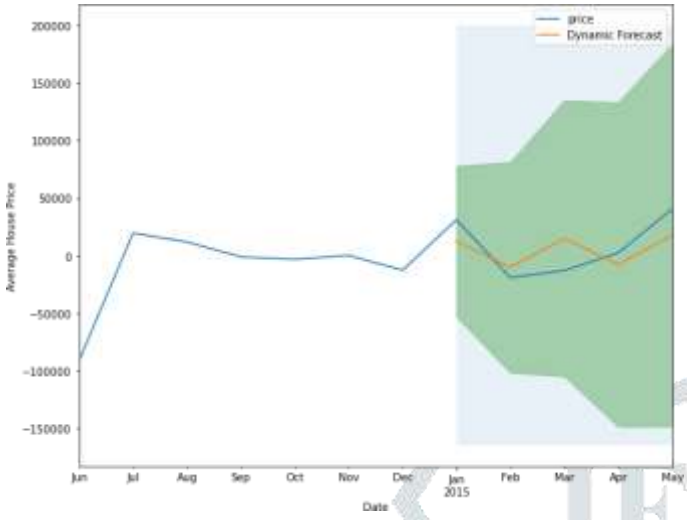

Fig. 2. SARIMA Model.

3) Content Based Recommendation System: Recommendation system is a machine learning system that gives generalized recommendation to its users based on some data or using users' preferences. It produces a list of recommendations in any of the two ways:

- Collaborative filtering approaches build a model from the users past behavior (i.e., items purchased or searched by the user) as well as similar decisions made by other users.
- Content-based filtering uses a series of discrete characteristics of an item in order to recommend additional items with similar properties.

We are going to use content-based filtering methods in our project. For example, in our project data set contains tuples like place, square feet, number of bhk , flat is furnished or not and floor number at which given flat is situated. Suppose client wants a new two bhk fully furnished flat in a area like diva of 300 sqrt feet at a floor maximum up to 5. After that recommendation system takes this and makes some assumptions client has entered floor number is 5 than system will search the flat for floor number 4 to 6 from training dataset that means if product is not available then it will try to give maximum similar type of product. Our will match every preference made by the user with the values present in the training data set and will try to give similar type of property.

C. Datasets Used

We are using 'kp house data' dataset for house prediction. This dataset is extracted by using web scraping concept. This dataset includes zipcode of different cities and attributes of house. Ex. sq.ft, bedrooms,floors,etc.

It contains 21614 record of various properties which are used for prediction and we are using same dataset for forecasting also.

D. Backend Connectivity

- This web application is connected to backend using python flask framework.
- Flask-MySQL is a Flask extension which allows to access a MySQL database.

IV. Execution and Output

A. Catboost Regressor

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	CatBoost Regressor	77461.8963	17741186840.8751	133196.0486	0.8692	0.1902	0.1479

Fig. 3. CatBoost Regressor.

B. Prediction Output for 9 months

	Zipcode	Current Value	1 Months Value	2 Months Value	3 Months Value	4 Months Value	5 Months Value	6 Months Val
0	98001	315466.541667	368575.531250	397427.322917	379069.762731	384705.918595	386219.114583	357167.3090
1	98002	229610.900000	276713.466667	280254.107143	286322.424242	286460.000000	267339.871795	261386.7500
2	98003	293499.736842	294067.649573	313827.280500	323425.149573	280566.673382	259052.716239	295307.3717
3	98004	1613529.273810	2068456.151261	1785036.470588	1900742.516340	1659151.484594	1818527.680083	1688745.2941
4	98005	930785.714286	855010.596537	926770.014286	885768.442857	936528.020779	937359.712698	905369.0904
5	98006	850093.043478	1753815.759425	1721487.477679	1636189.382440	1699260.868238	1737507.176091	1637961.7395
6	98007	589856.785714	606391.666667	590671.212121	471477.111111	501945.000000	558514.785714	421043.7500
7	98008	763521.759259	627523.194444	494465.587607	425699.011438	492560.347222	438118.694444	395268.1517
8	98010	440122.875000	433413.888889	488862.500000	402921.428571	362977.500000	487210.714286	425200.0000
9	98011	520772.727273	503952.222222	523183.385417	447552.666667	511599.444444	549742.435897	491501.4102

Fig. 4. Predicted Values.

C. Predicted Price

price	Label
285000.000000	328399.356569
239950.000000	236320.781714
460000.000000	474212.422349
397500.000000	498626.206444
545000.000000	598658.175695
...	...
455000.000000	484459.510900
850000.000000	742794.857716

Fig. 5. Predicted Values.

V. CONCLUSION

- In this project we studied that there is a great opportunities in the real estate market.
- House Price Prediction is very useful to make investments in the real estate.
- This system will help people get financial output much higher than their investment.
- The main objective of using this prediction and recommendation system is to reduce the human physical calculation, time and carry out the whole process at ease.

VI. REFERENCES

1) Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita. House Price Prediction using various Regression Analysis and Particle Swarm Optimization. (IJACSA) International Journal of Advanced Computer Science and Applica- tions, Vol. 8, No. 10, 2017

2) CH.Raga Madhuri, Anuradha G, M.Vani Pujitha. House Price Predic- tion Using Regression Techniques: A Comparative Study. IEEE International Conference on smart structures and systems ICSSS 2019.

3) Parasich Andrey Viktorovich, Parasich Viktor Aleksandrovich, Kaftan- nikov Igor Leopoldovich, Parasich Irina Vasilevna. Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning.

4) CH.Raga Madhuri, Anuradha G, M.Vani Pujitha. House Price Prediction Using Regression Techniques: A Comparative Study. IEEE International Conference on smart structures and systems ICSSS 2019.