House Price Prediction Using Machine Learning and RPA

Prof. Pradnya Patil

Assistant Professor, Computer Engineering Department, K. J. Somaiya Institute of Engineering and Information Technology, Mumbai, India

Darshil Shah, Harshad Rajput, Jay Chheda

Undergraduate Student, Computer Engineering Department, K. J. Somaiya Institute of Engineering and Information Technology, Mumbai, India

---***---

***Abstract*—**In today's world, everyone wishes for a house that suits their lifestyle and provides amenities according to their needs. House prices keep on changing very frequently which proves that house prices are often exaggerated. There are many factors that have to be taken into consideration for predicting house prices such as location, number of rooms, carpet area, how old the property is? and other basic local amenities. We will be using CatBoost algorithm along with Robotic Process Automation for real-time data extraction. Robotic Process Automation involves the use of software robots to automate the tasks of data extraction while machine learning algorithm is used to predict house prices with respect to the dataset.

***Keywords*—Random Forest, CAT Boost, RPA, House Price Prediction.**

I. Introduction

Data is the heart of machine learning. Predictive models use data for training which gives somewhat accurate results. Without data we can't train the model. Machine learning involves building these models from data and uses them to predict new data. Machine Learning is a subset of Artificial Intelligence. It gives system capability to learn wherein it automatically learns and improves its performance without being explicitly programmed. It does focus on the development of programs and use it to learn for themselves. As the world is moving forward to using variants technologies, so has automation improved its ways to make our work easier. Though the word automation was coined in the 1950s, very few people really understood what it meant. Robotic process automation means automating operations on business by using software robots to reduce human efforts. Robotics are entities that mimic human actions are called Robots. A process is a sequence of steps that leads to meaningful activity. For example, the process of making a dish or the process of merging two or more things into one.

Any process that is carried out by a robot without humans interfering in it is called Automation. Machine learning is closely related to statistics, which focuses on making predictions using computers. There are a variety of applications of Machine Learning such as filtering of emails, where it is difficult to develop a conventional algorithm to perform the task effectively. Machine learning algorithms are purely based on data. Machine Learning algorithms are an advanced version of the regular algorithm. It makes programs "smarter" by allowing them to automatically learn from the data provided by us. The algorithm is mainly divided into two phases and that is the training phase and the testing phase. Broadly there are three types of algorithms that are mainly used on data and they are supervised, unsupervised and reinforcement learning algorithms.

In Supervised learning, the algorithm consists of a target variable or a dependent variable which is to be predicted from a set of independent variables. Using a function, the inputs are mapped to the desired outputs. Examples of Supervised learning algorithms are Regression, Decision Tree, Random Forest, KNN, Logistic Regression, etc.

In Unsupervised learning, the algorithm does not have any target variable. It is used for clustering into different groups. Apriori algorithm, K-means, Principal Component Analysis, Independent Component Analysis are some examples of Unsupervised learning algorithms.

When the machine is used to make specific decisions, Reinforcement Learning is used. In this, the model is in an environment where it trains itself making it more accurate by using the trial and error methodology. The model hence learns from past experiences and it captures the knowledge about that domain to make accurate decisions, Example of Reinforcement Learning: Markov Decision Process-hot encoding is one such Reinforcement learning algorithm.

II. Literature Survey

Trends in housing prices indicate the current economic situation and also are a concern to the buyers and sellers. There are many factors that have an impact on house prices, such as the number of bedrooms and bathrooms. House price depends upon its location as well. A house with great accessibility to highways, schools, malls, employment opportunities, would have a greater price as compared to a house with no such accessibility. Predicting house prices manually is a difficult task and generally not very accurate, hence there are many systems developed for house price prediction. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh [1] had proposed an advanced house prediction system using linear regression. This system's aim was to make a model that can give us a good house price prediction based on other variables. They used the Linear Regression for Ames dataset and hence it gave good accuracy. The house price prediction project had two modules namely, Admin and the User. Admin can add location and view the location. Admin had the authority to add density on the basis of per unit area. Users can view the location and see the predicted housing price for that particular location.

This paper [1] proposed on Hybrid Regression technique for housing Prices Prediction focused on the use of creative feature engineering to find the optimal features and their correlation with Sales Prices. Feature engineering improved the data normality and linearity of data. Their system showed that working on the Ames Housing dataset was convenient and showed that the use of Hybrid algorithms (65% Lasso and 35% Gradient Boost) provided results in predicting the house prices rather than using one from lasso, ridge or gradient boost.

The paper proposed by Ayush Varma Abhijit Sharma Sagar Doshi Rohini Nair[2] suggested that the use of neural networks along with linear and boosted algorithms improved prediction accuracy. The dataset used here contained various essential parameters.The dataset was cleaned up. Three algorithms were used namely Linear Regression, Forest Regression and Boosted Regression. The dataset was tested on all three and the results of all the above algorithms were fed as an input to the neural network. Neural networks were used mainly to compare all the predictions and display the most accurate result. A neural network along with Boosted Regression was used to increase the accuracy of the result.

The paper proposed by Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy[3] shows the prediction model is based on regression analysis and particle swarm optimization(PSO). Hedonic pricing is implemented using regression techniques to predict the NJOP price(Dependent Variable) in the city of Malang, based on factors such as land area, NJOP land price, NJOP building price. PSO is a stochastic optimization technique used for the selection of affect variables. The results obtained show a minimum prediction error RMSE of 14.186.

This paper[4] surveyed to predict a continuous target value, using algorithms Logistic Regression, Support Vector Machine, Lasso Regression Technique, Decision Tree are used to build a predictive model. They have used a stepwise approach from Data Collection, Data Processing, Data Analysis, to Evaluating Models. Then the predicted output is stored in a CSV file. It was found that the Decision Tree had the best accuracy of 84% approx., they tried to implement the problem of Regression using the Classification Algorithm which was successful. They had used predefined open-source Kaggle Dataset consisting of 80 parameters, from which 37 parameters were chosen which were affecting house prices.

III. Proposed System

The world is shifting from manual to automated systems. The objective of our project is to reduce the problems faced by the customer. In the present situation, the customer visits a real estate agent so that he/she can suggest suitable showplaces for his investments. But the above method is risky as the agent may forecast wrong prices to the customer and that will lead to loss of customer's investment. This manual technique which is currently used in the market is outdated and has a high risk. So as to overcome the drawback, there is a need for an updated and automated system.

In our proposed system, the initial step is data scraping. It is a technique with the help of which structured data can be extracted from the web or any application and saved to a database or spreadsheet or CSV file. We will be using the UIPath Studio Platform to develop our RPA Flowchart. UiPath studio also provides the power data scraping with the assistance of scraping wizards.

After Data Extraction, we perform Data Cleaning. It refers to the modifications applied to the data before feeding it to the algorithm. Data Cleaning is a technique that is used to convert the raw data into a clean data set wherein we deal with missing data, categorical data as per the required needs. We have cleaned up our entire dataset and also truncated the outlier values.

After completion of Cleaning, we will apply various algorithms. There are many algorithms that can be used to predict the house rate. XGBoost, Light GBM, CatBoost are some of the algorithms that can be used. We will be using these algorithms for the prediction:

Random Forest is a trademarked term for an ensemble of decision trees. In Random Forest, we have many decision trees. Each tree gives a classification to classify a new object based on the attributes which mean that the tree votes for that class. The forest chooses the classification having the most votes (over all the trees within the forest).

In short, with Random Forest we can train the model efficiently for small amounts of data and can get pretty good results. It will, however quickly reach some extent where more samples won't improve the accuracy.

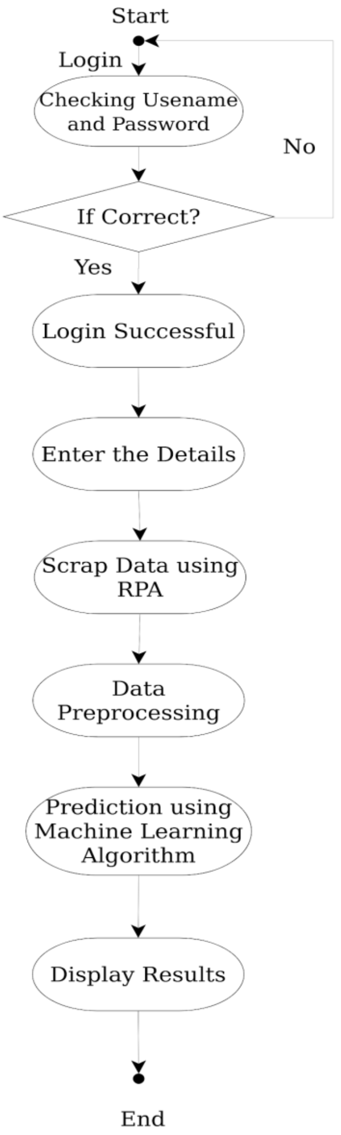
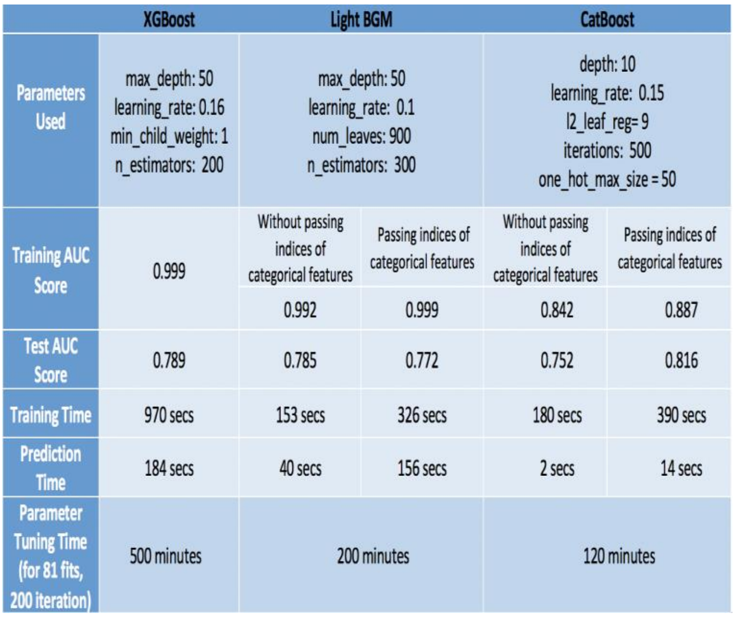


Figure 3.1

CatBoost is an open-source machine learning algorithm from Yandex. It can easily integrate with other deep learning frameworks if needed. CatBoost can work on a large variety of data and it does not require extensive data training. Catboost can automatically deal with categorical variables without showing the type conversion error unlike other algorithms such as XGBoost and Light GM. It helps you to focus on tuning your model better rather than sorting out trivial errors. Missing values are processed using two modes and they are "Min" and "Max". The minimum value for a feature is processed as the missing value. The missing value is given a value that is less than all the values. This way, it is guaranteed that a split that separates missing values from all other values is considered when selecting splits. In "Max", the maximum values among all the values are processed as a missing value.

After analysis of boosting algorithms based on a dataset of flight delays which contains both categorical as well as numerical features, suggests that CatBoost is a clear winner in terms of accuracy as shown in the table. The proposed system will be easy to use and will contain simple operations as shown in Figure 3.1.



	XGBoost	Light BGM		CatBoost	
Parameters Used	max_depth: 50 learning_rate: 0.16 min_child_weight: 1 n_estimators: 200	max_depth: 50 learning_rate: 0.1 num_leaves: 900 n_estimators: 300		depth: 10 learning_rate: 0.15 l2_leaf_reg= 9 iterations: 500 one_hot_max_size = 50	
Training AUC Score	0.999	Without passing indices of categorical features	Passing indices of categorical features	Without passing indices of categorical features	Passing indices of categorical features
		0.992	0.999	0.842	0.887
Test AUC Score	0.789	0.785	0.772	0.752	0.816
Training Time	970 secs	153 secs	326 secs	180 secs	390 secs
Prediction Time	184 secs	40 secs	156 secs	2 secs	14 secs
Parameter Tuning Time (for 81 fits, 200 iteration)	500 minutes	200 minutes		120 minutes	

IV. Conclusion

The various benefits that RPA provides to the market have made it one of the top contenders in the current market as the field of interest of many organizations worldwide. Most of the organizations are already implementing RPA technology as it generates more accurate and consistent processes that are less prone to errors. The system uses RPA to extract the data and also makes optimal use of machine learning algorithms which satisfies the customer by providing accurate output and preventing the risk of investing in the wrong house.

V.References

[1] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, "A Hybrid Regression Technique for House Price Prediction", December 2017.

[2] Ayush Varma, Abhijit Sharma, Sagar Doshi, Rohini Nair, "House Price Prediction Using Machine Learning And Neural Networks", INSPEC number 18116205, April 2018.

[3] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017.

[4] Neelam Shinde, Kiran Gawande, "Valuation of House Price Using Predictive Techniques", International

Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835(IJAECS), Volume-5, Issue-6, June-2018.

[5] Eli Beracha, Ben T Gilbert, Tyler Kjorstad, Kiplan womack, "On the Relation between Local Amenities and House Price Dynamics", Journal of Real estate Economics,Aug. 2016.

[6] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?” In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7376 LNAI, 2012, pp. 154–168,ISBN: 9783642315367. DOI: 10 . 1007 / 978 - 3 - 642 -31537-4\ 13.

[7] S. Ray, “CatBoost: A machine learning library to handle categorical (CAT) data automatically,” CatBoost: Analytics Vidhya, 14-Aug-2017.

[8] R. J. Shiller, “Understanding recent trends in house prices and home ownership,” National Bureau of Economic Research, Working Paper 13553, Oct. 2007. DOI: 10.3386/w13553.

[9] S. C. Bourassa, E. Cantoni, and M. Hoesli, “Predicting house prices with spatial dependence: a comparison of alternative methods,” Journal of Real Estate Research, vol. 32, no. 2, pp.139–160, 2010.

[10] Li, Li, and Kai-Hsuan Chu. "Prediction of real estate price variation based on economic parameters." Applied System Innovation (ICASI), 2017 International Conference on.IEEE, 2017.

[11] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011): 2825-2830.

[12] Byeonghwa Park , Jae Kwon Bae (2015). Using machine learning algorithms for housing price prediction , Volume 42, Pages 2928-2934.

[13] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, 2015. Introduction to Linear Regression Analysis.