Housing Price Prediction using Machine Learning

Kummaguri Jeevitha

Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *Machine learning plays a major role from past years in image detection, spam reorganization, normal speech command, product recommendation and medical diagnosis. Present machine learning algorithm helps us in enhancing security alerts, ensuring public safety and improve medical enhancements. Machine learning system also provides better customer service and safer automobile systems. In the present paper we discuss about the prediction of future housing prices that is generated by machine learning algorithm. For the selection of prediction methods we compare and explore various prediction methods. We utilize lasso regression as our model because of its adaptable and probabilistic methodology on model selection. Our result exhibit that our approach of the issue need to be successful, and has the ability to process predictions that would be comparative with other house cost prediction models. More over on other hand housing value indices, the advancement of a housing cost prediction that tend to the advancement of real estate policies schemes. This study utilizes machine learning algorithms as a research method that develops housing price prediction models. We create a housing cost prediction model In view of machine learning algorithm models for example, XGBoost, lasso regression and neural system on look at their order precision execution. We in that point recommend a housing cost prediction model to support a house vender or a real estate agent for better information based on the valuation of house. Those examinations exhibit that lasso regression algorithm, in view of accuracy, reliably outperforms alternate models in the execution of housing cost prediction.*

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this paper we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to particular person will be safe or not.

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.

1.1 Artificial Intelligence:

Artificial intelligence (AI) is the ability of a computer program or a machine to think and learn. It is also a field of study which tries to make computers "smart". As machines become increasingly capable, mental facilities once thought to require intelligence are removed from the definition. AI is an area of computer sciences that emphasizes the creation of intelligent

machines that work and reacts like humans. Some of the activities computers with artificial intelligence are designed for include: Face recognition, Learning, Planning, Decision making etc.,

Artificial intelligence is the use of computer science programming to imitate human thought and action by analysing data and surroundings, solving or anticipating problems and learning or self-teaching to adapt to a variety of tasks.

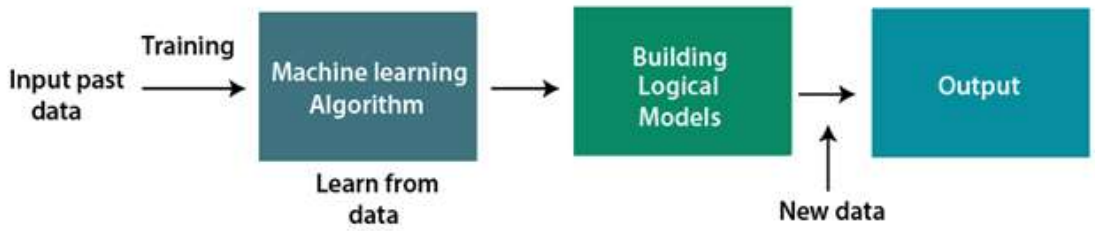
1.2 Machine Learning

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information**. Currently, it is being used for various tasks such as **image recognition**, **speech recognition**, **email filtering**, **Facebook auto-tagging**, **recommender system**, and many more.

Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by **Arthur Samuel** in **1959**. We can define it in a summarized way as: "Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed".

A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:



1.2.1 Features of Machine Learning:

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

1.2.2 Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

1) Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

Supervised learning can be grouped further in two categories of algorithms:

- **Classification**
- **Regression**

2) Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data.

It can be further classifieds into two categories of algorithms:

- **Clustering**
- **Association**

1.3 Neural Network

Artificial Neural Networks (ANNs) make up an integral part of the Deep Learning process. They are inspired by the neurological structure of the human brain. ANNs are "complex computer code written with the number of simple, highly interconnected processing elements which is inspired by human biological brain structure for simulating human brain working & processing data (Information) models."

Deep Learning focuses on five core Neural Networks, including:

- Multi-Layer Perceptron
- Radial Basis Network
- Recurrent Neural Networks
- Generative Adversarial Networks
- Convolutional Neural Networks.

II. LITERATURE SURVEY

[1] TITLE: House Resale Price Prediction Using Classification Algorithms

AUTHORS: P. Durganjali; M. Vani Pujitha

DESCRIPTION: Now a days house resale is majorly seen in metro cities. The market demand for housing is always increasing every year due to increase in population and migrating to other cities for their financial purpose. Prediction of house resale price for long-term temporary basis is important especially for the people who stays who will stay the long time period but not permanent and the people who do not want to take any risk during the house construction. In this paper, the resale price prediction of the house is done using different classification algorithms like Logistic regression, Decision tree, Naive Bayes and Random forest is used and we use AdaBoost algorithm for boosting up the weak learners to strong learners. Several factors that are affecting the house resale price includes the physical attributes, location as well as several economic factors persuading at that time. Here we consider accuracy as the performance metrics for different datasets and these algorithms are applied and compared to discover the most appropriate method that can be used the reference for determining the resale price by the sellers.

[2] TITLE: Housing Price Mathematical Prediction Method through Big Data Analysis and Improved Linear Regression Model

AUTHORS: Xinyu Yang; Zesheng Yin; Jiayi Li

DESCRIPTION: Housing price prediction is a typical case of exploratory data analysis. This article takes housing prices in a certain city in the United States as an example, and uses a variety of linear models to predict and analyze the trend of housing prices. Based on Spark, we have implemented a distributed housing price prediction analysis model. Through the comparison of multiple linear models, we find out the best linear analysis model with the best prediction results.

[3] TITLE: Housing prices prediction with deep learning: an application for the real estate market in Taiwan

AUTHORS: Choujun Zhan; Zeqiong Wu; Yonglin Liu;

DESCRIPTION: The housing market is increasing huge, predicting housing prices is not only important for a business issue, but also for people. However, housing price fluctuations have a lot of influencing factors. Also, there is a non-linear relationship between housing prices and housing factors. Most econometric or statistical models cannot capture non-linear relationships yet. Therefore, we propose housing price prediction models based on deep learning methods, which can capture non-linear relationships. In this work, we construct a dataset, including the housing attributes data and macroeconomic data in Taiwan from January 2013 to December 2018. The housing attributes data includes two types of housing transactions, which are "land + building" (Type1) and "land + building + park" (Type2). Macroeconomic data includes housing investment demand ratio, owner-occupier housing ratio, housing price to income ratio, housing loan burden ratio, and housing bargaining space ratio. Then, this dataset is utilized to evaluate the prediction methods based on deep learning algorithms BPNN and CNN to predict housing prices. Experimental results show that CNN with housing features has the best prediction effect. This study can be used to develop targetted interventions aimed at the housing market.

[4] TITLE: Housing Price Prediction Based on CNN

AUTHORS: Yong Piao; Ansheng Chen; Zhendong Shang

DESCRIPTION: Housing price has been one of the most concerned issues to the public all over the world. The excessive growth of housing price will affect not merely the quality of life, but also the business cycle dynamics. However, the factors influencing residential real estate prices are complex and the selection of effective features is vague, which leads to a lower accuracy in many of the traditional housing price prediction approaches. Accordingly, a novel prediction model based on CNN is proposed for prediction of housing price as well as the process of feature selection. Compared with other traditional methods, our work can obtain a better performance through experiments using actual data of property transaction.

III. EXISTING SYSTEM

- House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets.
- Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of Melbourne.
- Moreover, experiments demonstrate that the combination of Stepwise and Support Vector Machine that is based on mean squared error measurement is a competitive approach.

Disadvantages

- Less Accurate
- Data analysis is not proper
- Prediction is inaccurate.

IV. PROPOSED SYSTEM

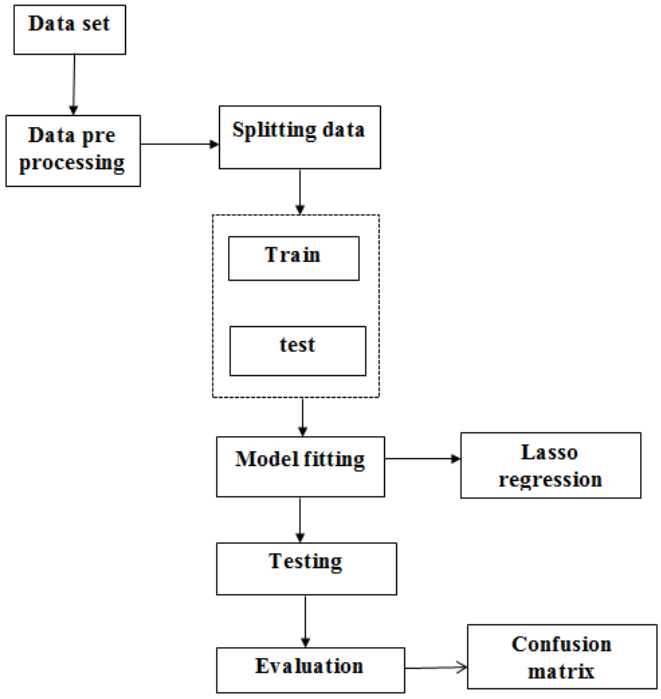
- We propose an approach for house price detection using machine learning techniques.
- The proposed approach uses Lasso Regression and Gradient Boost algorithm for model fitting.

- lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable.
- Gradient boosting It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error.

Advantages

- High performance and accuracy calculate.
- Data flow is passed well.
- Time Consuming.

V. SYSTEM ARCHITECTURE



5.1 Module Description

The following are the detailed explanation of the various modules and its method of execution. There are four modules in the system,

- ✓ Module 1: Data collection
- ✓ Module 2: Pre-Processing
- ✓ Module 3: EDA
- ✓ Module 4: Model Fitting
- ✓ Module 5: Evaluation

MODULE 1: Data collection

We are going to work with the house price dataset that contains various features and information about the house and its sale price. Using the *'read_csv'* function provided by the Pandas package, we can import the data into our python environment. After importing the data, we can use the 'head' function to get a glimpse of our dataset.

MODULE 2: Pre-Processing

The sklearn.preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.

In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the set, robust scalers or transformers are more appropriate. The behaviors of the different scalers, transformers, and normalizers on a dataset containing marginal outliers is highlighted in Compare the effect of different scalers on data with outliers.

MODULE 3: EDA

Exploratory Data analysis (EDA) is used for visualize the datasets

To visualize the dataset like pie chart, bar chart, box plot, histogram graph etc.,

MODULE 4: Model Fitting

In this proposed system we are using five machine learning algorithms named as Lasso Regression and Gradient Boosting

We could able to train the system using these four algorithms and evaluate training score calculated.

To predict the House price amount

MODULE 5: Evaluation

The system will predict the soil type with 98% accuracy

- To evaluate the accuracy score using confusion matrix method

  A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

Confusion matrix include:

- Precision
- Recall
- Support
- F1score
- Accuracy

VI. 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.

REFERENCES

[1] 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.