HOUSE PRICE PREDICTION USING MACHINE LEARNING

Name	R GOVINTHAN
Reg.No	950621104028
Date	28 SEPTEMBER 2023
Team ID	PROJ_212169_TEAM_1
Project Name	HOUSE PRICE PREDICTION USING ML
Maximum Marks	

Three Tasks:

1. **PROBLEM DEFINITION**

2. **DESIGN THINKING**

3. **PROBLEM SOLUTION**

1. DEFINITION

The trend of the sudden drop or constant rising of housing prices has attracted interest from the researcher as well as many other interested people. There have been various research works that use different methods and techniques to address the question of the changing of house prices. This work considers the issue of changing house price as a classification problem and discuss machine learning techniques to predict whether house prices will rise or fall using available data. This work applies various feature selection techniques such as variance influence factor, Information value, principle component analysis, and data transformation techniques such as outlier and missing value treatment as well as different transformation techniques. The performance of the machine learning techniques is measured by the four parameters of accuracy, precision, specificity, and sensitivity. The work considers two discrete values 0 and 1 as respective classes. If the value of the class is 0 then we consider that the price of the house has decreased and if the value of the class is 1 then we consider that the price of the house has increased.

2. DESIGN THINKING

Stage 1: Empathize - Research Your Users' Needs

Development of civilization is the foundation of the increase in demand for houses day by day. Accurate prediction of house prices has been always a fascination for buyers, sellers, and bankers also. Many researchers have already worked to unravel the mysteries of the prediction of house prices. Many theories have been given birth as a consequence of the research work contributed by various researchers all over the world. Some of these theories believe that the geographical location and culture of a particular area determine how the home prices will increase or decrease whereas other schools of thought emphasize the socio-economic conditions that largely play behind these house price rises.We all know that a house price is a number from some defined assortment, so obviously prediction of prices of houses is a regression task.

To forecast house prices one person usually tries to locate similar properties in his or her neighborhood and based on collected data that person will try to predict the house price.All these indicate that house price prediction is an emerging research area of regression that requires the knowledge of machine learning. This has motivated me to work in this domain.Realestate appraisal is an integral part of the property buying process. Traditionally, the appraisal is performed by professional appraisers specially trained for real estate valuation. For the buyers of real estate properties, an automated price estimation system can be useful to estimate the prices of properties currently on the market. Such a system can be particularly helpful for novice buyers who are buying a property for the first time, with little to no experience.

Stage 2: Define - State Your User’s Needs And Problems:

In India, there are multiple real estate classified websites where properties are listed for sell/buy/rent purposes such as 99acres, no broker, housing, magic bricks, and many more. However, in each of these websites, we can see a lot of inconsistencies in terms of pricing of a house and there are some cases when similar properties are priced differently and thus there is a lack of transparency and accuracy. Sometimes the customers may feel the value is not justified for a particular listed house but there is no way to confirm and check the data is accurate or not.Proper evaluations and justified prices of properties can bring in a lot of transparency and trust back to the real estate industry, which is very important as for most consumers, especially in India the transaction prices are quite high, and addressing this issue will help both the customers and the real estate industry in the long run.

We propose to use machine learning and artificial intelligence techniques to develop an algorithm that can predict housing prices based on certain input attributes. The business application of this algorithm is that classified websites can directly use this algorithm to predict prices of new properties that are going to be listed by taking some input variables and predicting the correct and justified price i.e. avoiding taking the wrong valuation for the house. This study is a proof-of-concept (POC) and can be treated as a valuation Report.

The purpose is to raise awareness about the correct valuation of property by accurate valuation. price inputs from customers and thus not letting any error creep into the system. This study on the proactive pricing of houses in the Indian context has never been reported earlier in the literature to the best of our knowledge. However, the problem of house price prediction is quite old and there have been many studies and competitions addressing the same including the Boston housing price challenge on Kaggle.

As far as housing price prediction in India is concerned, using machine learning techniques such as XGBoost for the prediction of housing prices in Bengaluru.MachineHack conducted a hackathon on predicting housing prices in Bengaluru in 2018. The problem statement was to predict the price of houses in Bengaluru given 9 features such as area type, availability, location, price, size, society, total square foot, number of bathrooms, and bedrooms. Moreover, there have been other studies for house price prediction in other cities of India such as Mumbai as well.

Stage 3: Ideate – Challenge Assumption And Create Ideas:

Machine Learning is a field of Artificial Intelligence that enables PC frameworks to learn and improve in execution with the assistance of information. It is used to study the construction of algorithms that make predictions on data. Machine learning is used to perform a lot of computing tasks. It is also used to make predictions with the use of computers. Machine learning is sometimes also used to devise complex models. The principle point of machine learning is to permit the PCs to learn things naturally without the assistance of people. Machine learning is very useful and is widely used around the whole world.

The process of machine learning involves providing data and then training the computers by building machine learning models with the help of various algorithms. Machine learning can be used to make various applications such as face detection applications, etc. Machine Learning is a field in software engineering that has changed the way of examining information colossally.

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.SVM Classifiers offer good accuracy and perform faster prediction compared to the Naïve Bayes algorithm. They also use less memory because they use a subset of training points in the decision phase. SVM works well with a clear margin of separation and with high dimensional space.

SVMs don't output probabilities natively, but probability calibration methods can be used to convert the output to class probabilities In the binary case, the probabilities are calibrated using Platt scaling: logistic regression on the SVM's scores, fit by additional cross-validation on the training data. So, here we will be using the machine learning technique of SVM to predict house prices by using various attributes to get the optimal and accurate house prices for the consumer.

Stage 4: Prototype – Start To Create Solution:

Stage 1: Collection of data

Information handling strategies and cycles are various. We gathered the information for Mumbai's land properties from different land sites. The information would have traits, for example, Location, cover region, developed region, age of the property, postal district, and so forth We should gather the quantitative information which is organized and ordered. Information assortment is required before any sort of AI research is completed. Dataset legitimacy is an unquestionable requirement in any case it is a waste of time to break down the information.

Stage 2: Information preprocessing

Data preprocessing is the most well-known approach to cleaning our instructive file. There might be missing characteristics or irregularities in the dataset. Data cleansing can help with these issues. Expecting a variable to have a large number of missing attributes we drop persons characteristics else substitute them with the typical worth.

Stage 3: The model's education

We should train the model first since the data is separated into two modules: a Training set and a Test set. The objective variable isjoined by the readiness set. The layout of educational assortment is computed using a decision tree regressor. A backslide model is collected as a tree structure by the Decision tree.

Stage 4: Testing and Integrating with UI

The test dataset is fed into the pre-programmed model, and home expenses are predicted. The front end, which includes Flask in Python, is then designed using

the pre-arranged model.

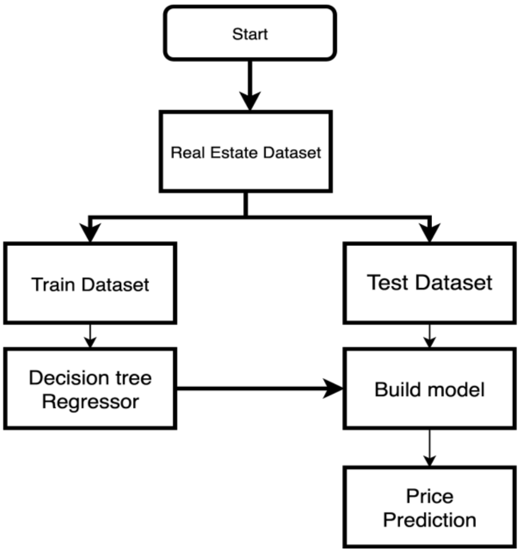


Figure 1 depicts the general development process.

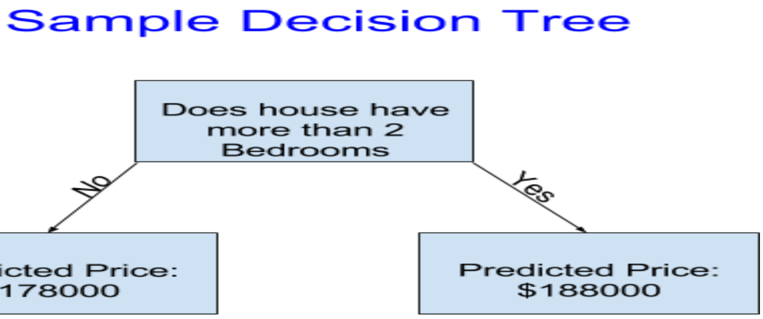
I. METHODOLOGY

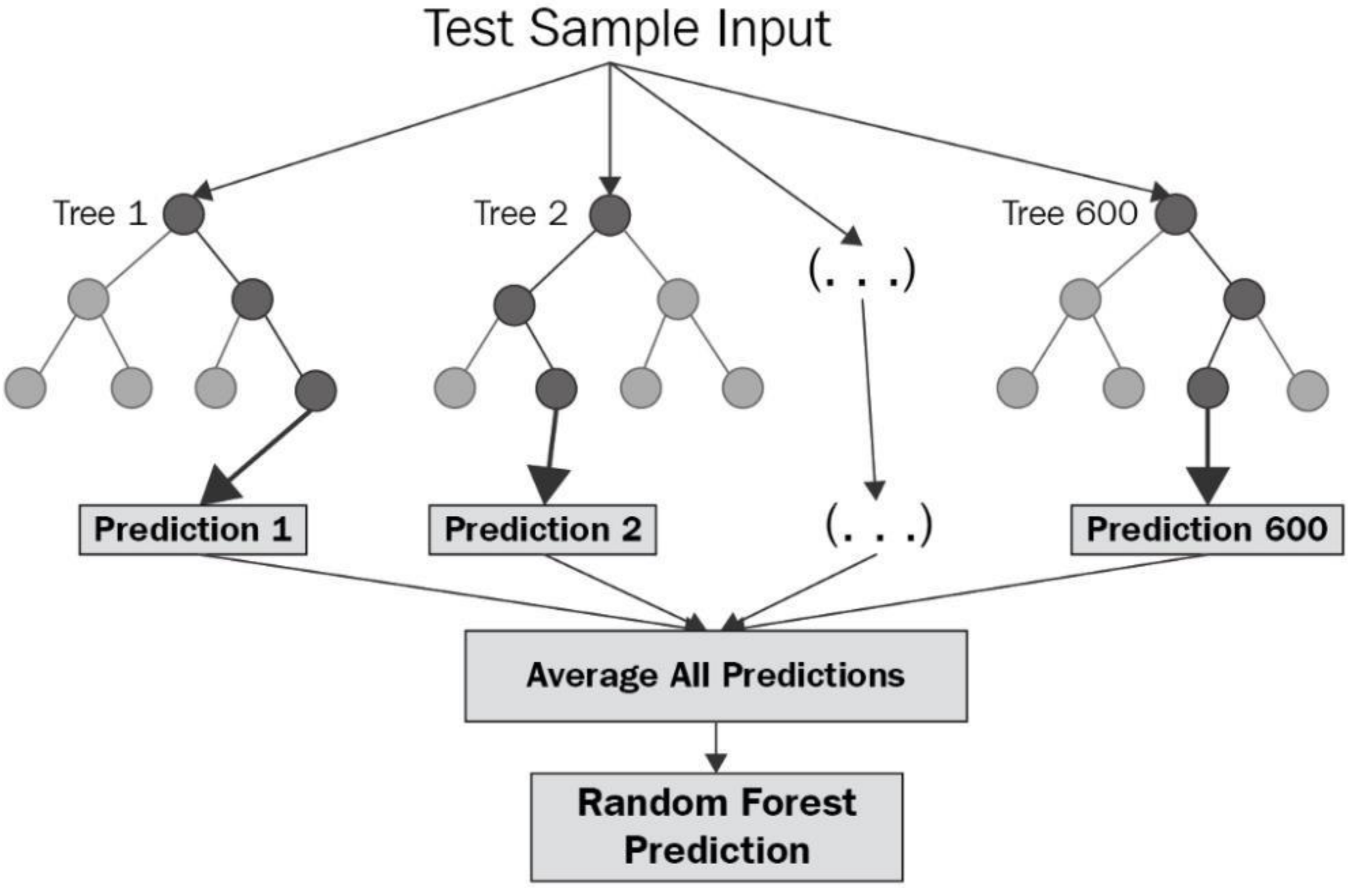
A. STUDIED ALGORITHMS:

During the time spent encouraging this model, different backslide computations were thought about. Straight backslide, Multiple straight backslide, Decision Tree Regressor, and KNN are all examples of machine learning techniques. were attempted upon the planning dataset. In any case, the decision tree regressor gave the most raised accuracy to the extent that expecting the house costs. The decision to pick the computation particularly depends on the angles and the type of data in the data that was used For our dataset, the decision tree computation is the best option.

B. REGRESSOR FOR DECISION TREE

The decision tree regressor recognises quality components and trains a model like a tree to forecast data in the future to provide a massive result. The highest significance and minimum significance of a chart are gained by the decision tree regressor, which then separates the data as demonstrated by the system Network Search CV is a strategy for overseeing limit tuning that will beneficially create and study a model for each mix of calculation limits exhibited in a cross-section. System Lookup In this calculation, CV is utilised to determine the optimal impetus for max- significance, which is then used to construct the decision tree.





Random Forest Structure

c. FLASK INTEGRATION

Right after building the model and giving the result, the accompanying stage is to do the consolidation with the UI, hence a cup is used. A carafe is a web structure. This implies that a carafe provides you with equipment, libraries, and headways that grant you to gather a web application. Flask is a framework for connecting Python models that is simple to learn.

II. IMPLEMENTATION

A. Data preprocessing:

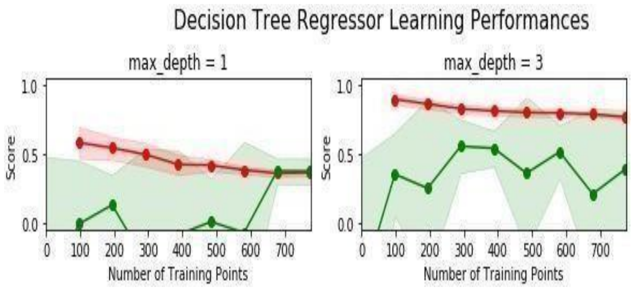
For their missing traits, age and floor restrictions were addressed. In addition, the preparation dataset's goal attribute is removed. This is why the Pandas library is used. The objective characteristic's min, max, standard deviation, and mean were identified for the factual representation of the dataset. We divided the dataset into two parts: a preparation set (80%) and a test set (20%). (20 percent).

B. Max-profundity:

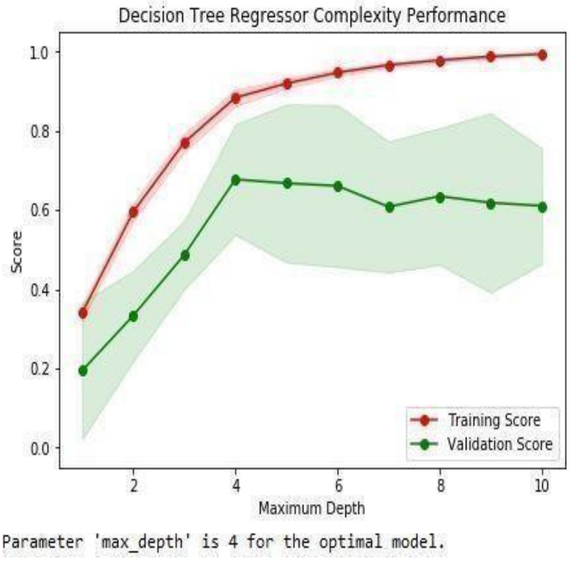
As previously said, system scan cv assists in determining the bush's maximum relevance. To visualise the various max-profunditiesand unpredictable execution, we utilised Matplotlib.

The visuals are as follows:

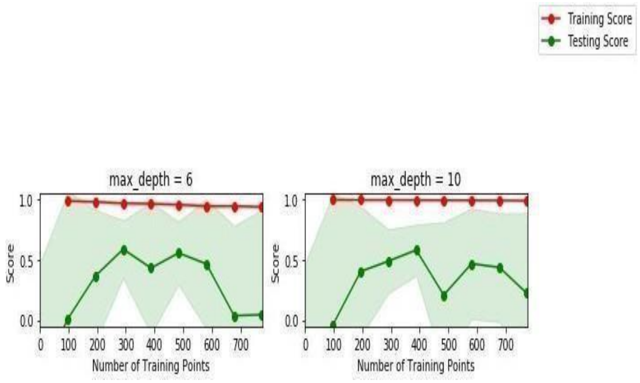
GRAPH



Maximal Profundity Values are Tested (1)(Score vs. Number of preparation points on pivot)



```
Parameter 'max_depth' is 4 for the optimal model.
```



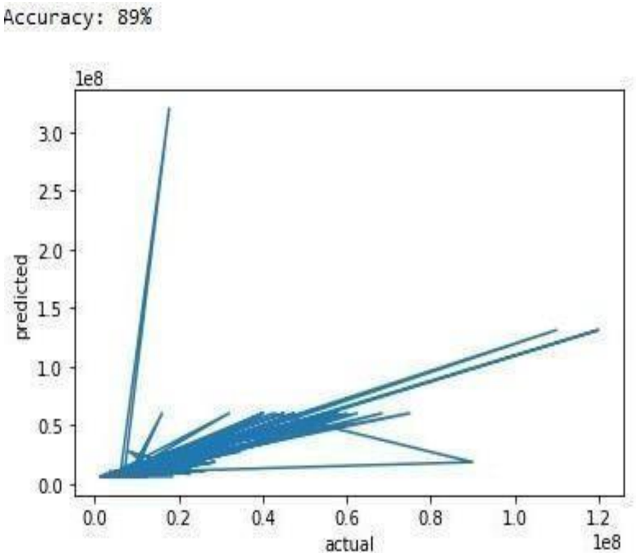
Max-profundity is an incentive for an ideal model

C. Fitting the model:

The model is built using a Decision tree regressor from the Scikit-learn module. The test set results are predicted using the predictwork.

III. OUTCOME

An accompanying displays the design of anticipated versus genuine costs with the precision of expectation:



```
Accuracy: 89%
```

Figure 5 shows a graph of actual vs. expected prices based on the dataset.

The r2 score of the regression model is what determines accuracy.

IV. FUTURE SCOPE

Future work on this study could be separated into four fundamental regions to further develop the outcome even further. This shouldbe possible by:-

- The utilized pre-handling strategies truly do help in the forecast exactness. Nonetheless, exploring different avenues regardingvarious blends of pre-handling strategies to accomplish better expectation exactness.

- Utilize the accessible elements and assuming they could be joined as binning highlights has shown that the information gotmoved along.
- Preparing the datasets with various relapse strategies, for example, Elastic net relapse that consolidates both L1 and L2standards. To grow the examination and check the execution.
- The connection has shown the relationship in the neighborhood information. In this way, endeavoring to improve the neighborhood information is expected to cause rich with highlights that fluctuate and can give a solid connection relationship.

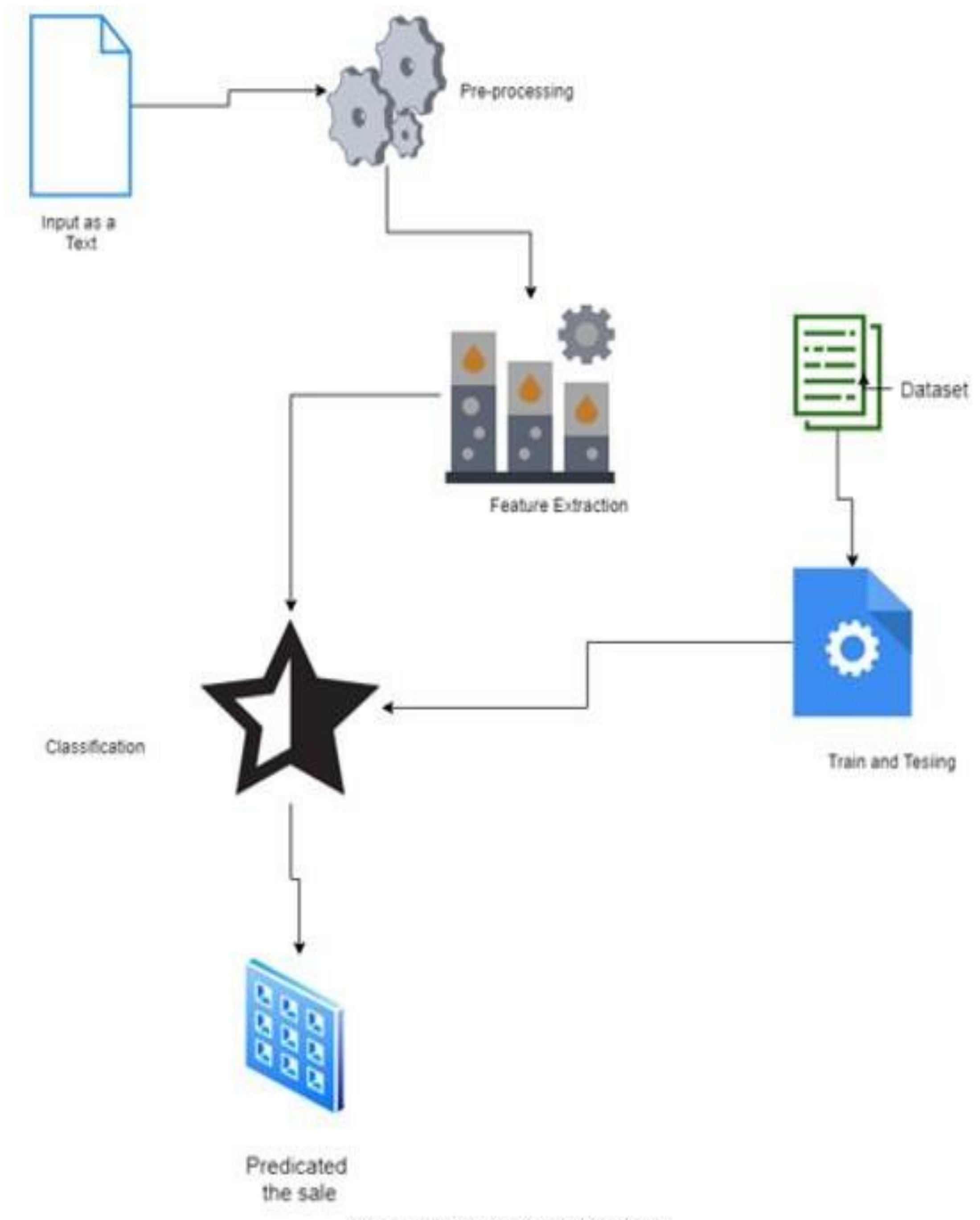


Fig 1. System Architecture

State 5:
Test :

The paper studies the SVM algorithm in machine learning for house price prediction. It takes data from the user and process it and classify it using pre-available data and uses various classification algorithm and classifies data and predict the accurate price of the property. It then confirms that accurate prediction result also depends on the population and quality of the training dataset. Results obtained earlier through SVM vs optimized SVM were then be evaluated. From the comparative analysis done in the next section, SVM shows comparable value over the other cryptocurrencies for this period. In the future, the model will be enhanced on the accuracy rate of the forecasted price. Future work will concentrate on the data preprocessing by including the sentiment data before the testing and training experiments.

3. PROBLEM SOLUTION

- Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. New analytical techniques of machine learning can be used in property research. This study is an exploratory attempt to use three machine learning algorithms in estimating housing prices, and then compare their results.

- In this study, our models are trained with 18-year of housing property data utilising stochastic gradient descent (SGD) based support vector regression (SVM), random forest (RF) and gradient boosting machine (GBM). We have demonstrated that advanced machine learning algorithms can achieve very accurate prediction of property prices, as evaluated by the performance metrics. Given our dataset used in this paper, our main conclusion is that RF and GBM are able to generate comparably accurate price estimations with lower prediction errors, compared with the SVM results.

- First, our study has shown that advanced machine learning algorithms like SVM, RF and GBM, are promising tools for property researchers to use in housing price predictions. However, we must be cautious that these machine learning tools also have their own limitations. There are often many potential features for researchers to choose and include in the models so that a very careful feature selection is essential.

- Second, many conventional estimation methods produce reasonably good estimates of the coefficients that unveil the relationship between output variable and predictor variables. These methods are intended to explain the real-world phenomena and to make predictions, respectively. They are used for developing and testing theories to perform causal explanation, prediction, and description (Shmuell, Citation2010). Based on these estimates, investigators can interpret the results and make policy recommendations. However, machine learning algorithms are often not developed to achieve these purposes. Although machine learning can produce model predictions with tremendously low errors, the estimated coefficients (or weights, in machine learning terminology) derived by the models may sometimes make it hard for interpretation.

- Third, the computation of machine learning algorithms often takes much longer time than conventional methods such as hedonic pricing model. The choice of algorithm depends on consideration of a number of factors such as the size of the data set, computing power of the equipment, and the availability of waiting time for the results. We recommend property valuers and researchers to use SVM for making forecasts if speed is a primary concern. When predictive accuracy is a key objective, RF and GBM should be considered instead.

- To conclude, the application of machine learning in property research is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to property appraisal, and presenting an alternative approach to the valuation of housing prices. Future direction of research may consider incorporating additional property transaction data from a larger geographical location with more features, or analysing other property types beyond housing development.