

House Price Prediction Taking Account of the Influence of COVID-19

Mingyu Bi

Faculty of Natural, Mathematics & Engineering Sciences, King's College London, London, England
k22003537@kcl.ac.uk

Abstract. Machine learning algorithms are utilized in various practical applications and academic research. Due to the advent of digital technology, there is now a greater availability of massive, organized, and georeferenced datasets. This has simplified the use of algorithms for trend analysis and detection as well as forecasting, which helps users make decisions. The purpose of this study is to determine which Machine Learning (ML) algorithms are most effective in projecting housing price and to assess how the COVID-19 has affected these prices. The method includes the stages of data preprocessing, feature manipulation, hyperparameter tuning, evaluation, selection, and, in the end, interpretation of the model. Another objective is to assess and compare several ensemble learning strategies, comprising bagging techniques like random forest, and boosting techniques like Gradient Boosting Regressor, using a linear regression model. The results of this study confirm that the real estate sector is very resistant to pandemics, as the decline in prices was not as severe as initially anticipated. Among the several models that were examined, extreme gradient boosting demonstrated the highest level of accuracy, albeit with a minimal difference.

Keywords: Machine Learning, Real Estate Market, COVID-19.

1. Introduction

Ever since the COVID-19 epidemic first appeared, this epidemic has had significant effects all over the world. Under the influence of economic dislocation, changes in people's lifestyle, and the introduction of new policies, the real estate industry has also experienced unprecedented changes. This can be revealed in several points. Firstly, the lockdown and restriction measures have led to a decrease in demand for the real estate market. Secondly, due to social distancing and economic uncertainty, purchasing and sales activities have also decreased. Lastly, COVID-19 has brought about turbulence in the financial market, which had a negative influence on the house price, investors may temporarily withdraw from real estate investment in this uncertain situation. Fluctuations in housing prices can be an essential part in driving economic growth or decline, the governments in different countries devised various measures to fight the negative consequences of the epidemic crisis on the house price [1].

As the influence made by the covid-19 in house price is uncontrollable, the prediction by machine learning is necessary due to their excellent feature extraction ability and prediction in many domains [2-4]. Algorithms for machine learning (ML) are increasingly being used in automated valuation models and large-scale real estate assessments. uniform techniques are employed in mass appraisals, ensuring that assessments are conducted in an impartial and standardized manner by gathering data from real estate offers and using it to estimate the values of big groups of properties [5-7]. These methodological technologies are very new, as the estimation of property prices has begun using machine learning techniques recently. Antipov et al. [8] evaluated 2695 properties in St. Petersburg, Russia, using machine learning and the random forest (RF) technique. Additionally, they used a variety of algorithms, including k-nearest neighbors (K-NN), Chi-squared automatic interaction detector (CHAID) and classification and regression tree (CART). A comprehensive assessment of prior research on the connection between TOM and pricing was done by Johnson et al. [9]. They discovered that around 20 studies found no discernible impact on price, 30 studies indicated a positive association, and 52 studies found a negative correlation. It is challenging to determine which covariates have the biggest influence on price corrections and in which direction because other

covariates in the literature also exhibit a wide range of reactions to real estate prices. Following a review of the literature, there is disagreement over the best algorithms or ML techniques for forecasting real estate values. There is general agreement that ML methods outperform classical linear models. Furthermore, authors like Renigier-Bilozor et al. [10] point out that new (automated) technologies must be used in the real estate appraisal field to supplement established approaches. For this reason, the objective is proposed in this research. Finding the best machine learning algorithms for a case study's house price prediction is the goal. This is how the rest of the paper is structured. This paper will first describe the tools and techniques, including the sources consulted and the database created in section 2. The debate is covered in Section 3, and Section 4 will provide a summary of the work and the conclusions.

2. Methods

2.1. Framework of AI-based Algorithms for House Price Prediction

Firstly, the data is collected by a web-scraping technique. Simply, it's an automated method of extracting structured data from websites. Secondly, it is the data pre-processing. One of the most crucial creative efforts in the data analysis was the retransformation of the unclear original data that was required for the data and future research. After that, it should be the training of data by using the ML testing model, such as linear regression, decision tree and lasso regression and so on. Finally, the evaluation of the results can find the best score of each testing model, to judge whether the model is precise to use. As the Fig. 1 shown:

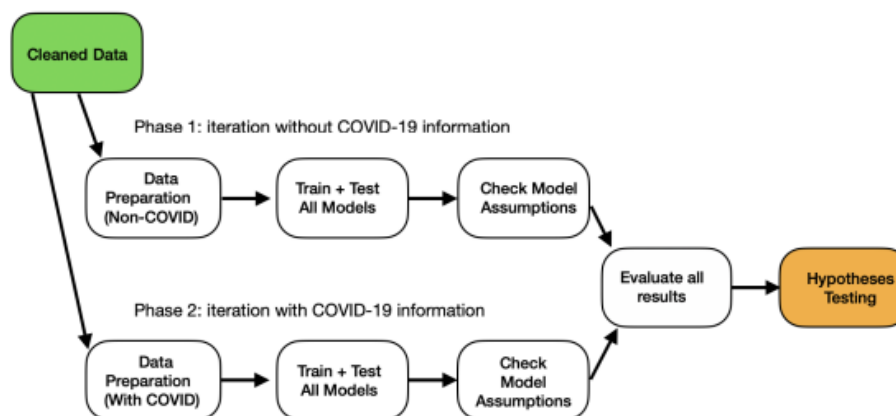


Figure 1. Experimental setup Phase 1 and 2 visualized [11]

2.2. ML Algorithm for House Price Prediction

In order to guarantee impartial interpretation while applying SHAP values, Grybauskas [11] split the dataset into more training and less test datasets. The most consistent machine learning method was then utilizing the training program in the interim months. The performance of the 15 models over a four-month period, encompassing both rental and sales operations, was then displayed based on several different criteria: accuracy, precision, and recall, and so on. Because of the dataset's imbalance and high proportion of negative cases, they concentrated on, F_1 score, precise percentages, and accuracy. In any case, higher ratios are preferable. The accuracy formula was as follows:

$$Accuracy = \frac{True\ positives + True\ negatives}{All\ samples} \quad (1)$$

Which provides the overall model accuracy because the denominator's samples were all used. In the meantime, the denominator's precision formula, which only included true positives and false positives, looked like this:

$$Precision = \frac{True\ positives}{True\ positives + False\ positives} \quad (2)$$

Furthermore, in the presence of unusual events, a different metric can integrate the compromises made between recall and precision to generate a single classifier measure. The F1 metric is called:

$$F1 = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \quad (3)$$

In summary, the F1 metrics, accuracy, and precision were the most crucial factors in determining the MCMLA.

2.3. Catboost Project for House Price Prediction During the Covid-19

Hu [12] used the Catboost project, a machine learning project they created. Datasets containing and devoid of pandemic period data yielded comparable results within the same model, suggesting that the pandemic had little effect on Perth's housing market. Simultaneously, the comparable results indicate that the pandemic is presumably not going to have a lasting impact either, which is dubious based on KPMG's conventional statistical approach research.

After analysis, it can be said that there isn't any clear proof that COVID-19 affects Perth real estate prices. In order to give machine learning models a chance in the experienced-based industry, such as real estate, a few risk analysis approaches in the VaR and ES perspectives were finally explored.

2.4. Boosting And Bagging Algorithms for House Price Prediction During the Covid-19

Mora-Garcia [13] contrasted ensemble learning methods according to bagging and boosting with a linear regression model. Additionally, a sizable database including aggregated cross-sectional data on home prices was produced using data from a variety of sources, including cadastral records, real estate portals, sociodemographic and economic variables, and satellite data. With the use of a sample of nearly 40,000 properties, a time period spanning from 2019 to 2021 was examined, enable to characterize the influence of COVID-19 on housing price. The result is although using machine learning algorithms is a multi-phase, intricate procedure, these algorithms outperform conventional linear methods. Given that the algorithms' efficacy fluctuates based on the situation at hand and the kind of data being utilized (text, picture, sound, and so on), it is impossible to generalize that one approach is superior to another. Furthermore, data leaking must be addressed in all ML research, and it must be determined whether overfitting of the employed algorithms could impair prediction accuracy.

2.5. Custom Web-scraping Algorithm for House Price Prediction During the Covid-19

Grybauskas [14] built a unique web scraping algorithm and gathered Vilnius real estate listings in the initial COVID-19 wave. After that, fifteen distinct machine learning models were used to predict apartment renovations. Each model was then assessed based on specific standards to determine which method was the most accurate. Moreover, this investigation was able to analyze the variable predictive power because of the latest advancements in SHAP values. The result is the pandemics don't seem to affect real estate too much, as evidenced by the less severe price declines than expected. For rent and sell activities, the four-month average price decrease was just – about 7% and 4%, respectively. Out of all the 15 models that were tested, the XGB model had the highest accuracy. Nevertheless, there was very little variance. The recovered SHAP showed that the most significant and reliable indicator for forecasting price revisions was the TOM variable. Moreover, the first pricing arrangement made logical. This suggests that there are two transition points for rent and sell activities, one at around 25 and 45 days and the other between 90–120 and 200–250 days.

3. Discussion

Adopting linearity over the whole distribution of the data is one of the primary issues with linear models. Applying machine learning techniques to replicate nonlinear patterns in various datasets, including the real estate market, solves this issue. This same conclusion has also been reached by studies like [15-18]. The boosting-based ensemble learning algorithms demonstrated superior

performance by using a variety of ways. They exhibit strong performance and are less susceptible to overfitting compared to other algorithms like random forest and extra trees. Lgbm and xgbm are the most effective models, in particular when utilizing big datasets. This is especially important for machine learning initiatives in the commercial sector in the stages of hyperparameter tuning and training. Strong performance has been shown by the bagging-based ensemble techniques, such as extra trees and random forest. However, they have a higher tendency to overfit the training data, potentially compromising their ability to generalize to unseen data. Overfitting occurs when a model has a high level of complexity, leading to a larger variation in its predictions [16]. The case study's model analysis indicates that the size of the floor, the quantity of bathrooms, and the existence of an elevator are crucial elements for precisely estimating property values. Nevertheless, there can be notable discrepancies based on distinct geographical areas and the specific timeframe under examination [19]. It is important to mention that location plays a significant role, particularly in a town along the sea as examined in this paper [17]. The study successfully detected the regions characterized by elevated pricing, as well as the regions with reduced costs situated in the northwestern part of the municipality. Since net family income directly affects consumers' purchasing power, it is a very reliable measure of real estate prices. Previous researches have acknowledged this association [17]. There are two difficulties in the experiment because of the influence of some unforeseen circumstances. Firstly, as mentioned before, governments in different countries have developed various measures to deal with the negative efforts of the pandemic crisis in the real estate market. By using the models, it's hard to predict the house price. Secondly, instead of falling, house prices in many countries increased dramatically in this catastrophic event [20], it means that There are many other factors that affect housing prices, so the prediction from the experiment can be no precise. To make the experiment more accurate, more and more factors that influence the house price are necessary in the model. Also, it's a good idea to find a new model to predict house prices, which can include more factors in the prediction.

4. Conclusion

To provide cutting-edge services and enhance industrial procedures, the great change in AI is being embraced by the real estate sector. This work evaluates the performance of multiple ensemble learning methods, such as boosting and bagging-based techniques, and juxtaposes them against a linear regression model. It could be inferred that, although it is a complex process, using machine learning algorithms is preferable to standard linear techniques. It's therefore hard to summarize that the strategy is superior to another because the efficiency of the algorithms depends on the current circumstances and the kinds of data like picture, sound, text and so on, will be employed. Moreover, it is crucial that any research including machine learning recognizes the issue of data leaks and evaluate whether the algorithms being used are overfitting could jeopardize the prediction accuracy. In the future, the further study plans to test the new assessment method to obtain more specific advice to improve the evaluation method.

References

- [1] Nicola M, Alsafi Z, Sohrabi C, Kerwan A, Al-Jabir A, Iosifidis C, Agha M, Agha R. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International Journal of Surgery*, 2020, 78: 185–93.
- [2] Gupta V, Joshi N, Mathur I. Advanced machine learning techniques in natural language processing for Indian languages. *Smart Techniques for a Smarter Planet: Towards Smarter Algorithms*. Cham: Springer International Publishing, 2019: 117-144.
- [3] Qiu Y, Wang J, Jin Z, et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. *Biomedical Signal Processing and Control*, 2022, 72: 103323.
- [4] Jones C, Castro D C, De Sousa Ribeiro F, et al. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 2024: 1-9.

- [5] Kauko T, d'Amato M. Introduction: Suitability Issues in Mass Appraisal Methodology. In *Mass Appraisal Methods*; Blackwell Publishing Ltd.: Oxford, UK, 2008; pp. 1–24.
- [6] Grover R. Mass valuations. *J. Prop. Investig. Financ.*, 2016, 34: 191–204.
- [7] IAAO, International Association of Assessing Officers. *Standard on Mass Appraisal of Real Property* (2017); International Association of Assessing Officers: Kansas City, MI, USA, 2019; p. 22.
- [8] Antipov EA, Pokryshevskaya EB. Mass appraisal of residential apartments: An application of Random Forest for valuation and a CART-based approach for model diagnostics. *Expert Syst. Appl.*, 2012, 39: 1772–1778.
- [9] Johnson K, Benefield J, Wiley J. The probability of sale for residential real estate. *J Hus Res.*, 2007, 16: 131–42.
- [10] Renigier-Biłozor M, Żróbek S, Walacik M, Janowski A. Hybridization of valuation procedures as a medicine supporting the real estate market and sustainable land use development during the covid-19 pandemic and afterwards. *Land Use Policy*, 2020, 99: 105070.
- [11] Grybauskas A, Pilinkienė V, Stundžienė A. Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. *J Big Data*, 2021, 8: 105.
- [12] Hu Y. *The Impact of COVID-19 on Perth House Price: A Machine Learning Perspective*. 2022.
- [13] Mora-Garcia R-T, Cespedes-Lopez M-F, Perez-Sanchez VR. Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. *Land*, 2022, 11: 2100.
- [14] Grybauskas A, Pilinkienė V, Stundžienė A. Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. *J Big Data*, 2021, 8: 105.
- [15] Hong J. An Application of XGBoost, LightGBM, CatBoost Algorithms on House Price Appraisal System. *Hous. Financ. Res.*, 2020, 4: 33–64.
- [16] Kok N, Koponen E-L, Martínez-Barbosa CA. Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *J. Portf. Manag.*, 2017, 43: 202–211.
- [17] Rico-Juan JR, Taltavull de La Paz P. Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Syst. Appl.*, 2021, 171: 114590.
- [18] De Aquino Afonso BK, Carvalho Melo L, Dihanster W, Sousa S, Berton L. Housing Prices Prediction with a Deep Learning and Random Forest Ensemble. In *Proceedings of the Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2019)*, Salvador de Bahia, Brazil, 15–18 October 2019; pp. 389–400.
- [19] McGreal WS, Taltavull de la Paz P. Implicit house prices: Variation over time and space in Spain. *Urban Stud.*, 2013, 50: 2024–2043.
- [20] Yiu CY. Why House Prices Increase in the COVID-19 Recession: A Five-Country Empirical Study on the Real Interest Rate Hypothesis. *Urban Sci.*, 2021, 5: 77.