House Price Prediction Using Neural Networks with Stacking Methods

Team 158
Chris Armstrong (carmstrong37)
Haoyang Han (hhan77)
Jianxing Niu (jniu37)
Pengfei Mei (pmei8)
Shiyu Du (sdu37)
Weiqian Pan (wpan48)

1. Summary

This project is based on open-source housing price data provided by Zillow. During the project, our team will develop a complete pipeline for data cleaning, descriptive analysis, visualization, feature engineering, and modeling. Traditionally, statistical transformation and analysis would be used to perform housing price prediction. However, more modern techniques, such as machine learning, multi-task learning, and deep learning (LSTM) can also be used. This project will use an ensemble/stacking model combining traditional and contemporary techniques (LR, kNN, RF, XGBoost, and MLP) for predicting house prices.

2. Background & Introduction

Housing price prediction plays a significant role in the modern real estate market.[8, 12,14] It is important for many private and public entities. Governments use house prices to determine interest rate policy, real estate buyers and sellers use prices to determine a "good" deal, and economic researchers use prices to judge the health of the economy. This project is to find a model (method) that can precisely predict house prices by combining multiple machine learning and deep learning algorithms.

Nowadays, many scholars research housing price prediction with various methods, like traditional time series models[3,13], classical machine learning algorithms[6, 12, 10], and classical deep learning algorithms[7,11, 17]. Researchers also have done a lot of work comparing different methods [6, 7, 17].

Reviewing all previous work, we have come to the following conclusions that will be used in our project.

- Location-based classification (zone) is critical for housing price, which sets the baseline of the unit price[7]
- 81.7% of total variation in house price can be explained by living area, number of the bedrooms & bathrooms, lot size, and age of house [9]
- 4% of total variation in house price will be affected by "soft conditions"[12], such as independent heating, good view, floor type, etc, which are not available in the dataset.

- There is no strong evidence for the role of macroeconomic fundamentals (unemployment rate, GDP growth, average income, etc) helping the model predict future home prices [8]
- Pure linear regression won't be accurate, more sophisticated machine learning or deep learning method is required [5,7,11,12,10,13]
- We should develop methodology from the following areas:
    - Machine learning algorithms, such as KNN, SVM, MLP, DT etc;
    - Deep learning algorithms, such as ANN, Black Box model, LSTM, SRN, etc;
    - Time series analysis methods, such as ARIMA, DFM, LBVAR etc, combined with autoregressive models.

We also find limitations in previous work. For example, time series analysis requires a large training dataset for historical housing price over a long time (~10 years or more), which is difficult to get. Thus we will just focus on static ("here and now") price estimation, i.e. estimate the house price within a certain "nearby" time window, rather than predicting house prices in the future. The DFM method seems to overestimate vibration at the price turning point. We will need to apply a smoothing method together with the DFM algorithm. Some of the paper referenced for our proposal only applies a single method in each prediction without trying to combine multiple algorithms to make a better prediction.

Based on the above, we propose our analysis methods and procedures.

3. Algorithm Design and Implementation

Designing a robust and accurate algorithm for house price prediction is a comprehensive process with lots of details to be considered. In this section, we will talk about the innovation part of the project; our desired impact; what we hope to achieve with our novel approach; risks and payoffs of the new approach; computational effectiveness; and how to evaluate the final outcome of project.

3.1 Innovation

Our literature review revealed that many comprehensive models have been implemented for house price prediction, such as bagging/boosting trees and stacked LSTM. We are trying to improve upon the previous work in 2 key ways:

- The first approach is innovative in the sense that we are applying techniques from Convolutional Neural Networks (CNN) to traditional Deep Neural Networks: employing various activation functions and a dropout layer. Another innovation we will try is the idea of skip-connection (allowing neurons in a lower layer to skip a few layers and connect to a much higher layer so the signals are easily learned there) from ResNet if this improves a deep model's performance. We believe we can achieve better performance with these innovations, and either combining multiple neural networks or combining the best neural networks with other models, such as SVM.
- The second approach is to develop a stacked algorithm, with kNN, XGBoost, LR, RF, and MLP in one model. Since ensemble method decision-trees can boost performance, we believe if we combine multiple algorithms, we can also boost performance.

3.2 Impact and Risk of the model

We will figure out the best model or combination of models that can predict house price accurately, leading to better confidence for buyers and sellers in the residential real estate market. Our model may be harder to understand intuitively and may require lots of computational resources. However the payoff comes in the form of buyers and sellers being able to worry less about inflated or deflated prices.

Our model may be harder to understand intuitively and may require lots of computational resources. However the payoff comes in the form of buyers and sellers being able to worry less about inflated or deflated prices. The traditional time series models use relatively less computer resources but ML/DL models can cost a lot, and the traditional regression model is highly explainable, but some ML models such as tree-based models are hard to explain. However, convolutional layer-based and stacking models have the potential of achieving

higher performance, which could reassure buyers and sellers of the predicted prices.

3.3 Cost of the Project

Because the team has access to CPUs and GPUs, and the dataset is publicly available, no cost is expected. The model should take less than 100MB in RAM, and be trained within a day or two. The prediction should be done within a second.

3.4 Evaluation Metrics

We will divide data into training, validation and test datasets. After testing the performance of the model by using multiple metrics for regression such as MSE, MAE, MAP, and RMSE, the midterm check passes if the prediction error is within 10% for the validation dataset on average, 15% for the test dataset on average. Final check passes if the prediction error is within 5% for validation dataset on average, 10% for the test dataset on average. Meanwhile, cross-validation metrics should also be minimized as much as possible.

4. Project Schedule

All team members have contributed equally. We propose the following project schedule:

Proposed Project Schedule		
Date	**Content**	**By**
March 4	Complete Proposal	Everyone
March 4	Complete Proposal presentation	Chris Armstrong
March 7	Gathering Dataset	Everyone
March 14	Data pre-processing and statistical analysis	Weiqian Pan
March 21	Developing model fitting	Everyone

	pipeline and benchmark	
March 29	Process Report	Everyone
April 5	Developing multiple model for comparison	Jianxing Niu, Haoyang Han, Pengfei Mei
April 12	Grid search/parameter tuning	Shiyu Du
April 15	Drawing Conclusion and generating report	Chris Armstrong
April 19	Final Report	Everyone
April 19	Final Presentation	Everyone

5. References

[1] Baldominos A., Blanco I., Moreno A. J., Iturrarte R., Bernárdez Ó., Afonso C. (2018 Nov 21). Identifying Real Estate Opportunities Using Machine Learning. Retrieved from https://arXiv:1809.04933v2

[2] Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M., Manjunath, B. S. (2017). Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). doi: 10.1109/wacv.2017.42

[3] Chen X., Wei L., Xu J. (2017 Sep 25). House price prediction using LSTM. Retrieved from https://arXiv:1709.08432

[4] Choudhary, P., Jain, A., & Baijal, R. (2018). Unravelling Airbnb Predicting Price for New Listing. Retrieved from https://arxiv.org/pdf/1805.12101.pdf

[5] Fan C., Cui Z., Zhong X. (2018 Feb) House Prices Prediction with Machine Learning Algorithms. Retrieved from http://delivery.acm.org/10.1145/3200000/3195133/p6-Fan.pdf?ip=123.126.70.237&id=3195133&acc=ACTIVE%20SERVICE&key=A79D83B43E50B5B8%2E5E2401E94B5C98E0%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1551414804_28fb84672ad92bbdbc8277634f1421f4

[6] Gao G., Bao Z., Cao J., Qin A. K., Sellis T., Wu Z. (2019 Jan 7). Location-Centered House Price Prediction: A Multi-Task Learning Approach. Retrieved from https://arxiv.org/abs/1901.01774

[7] Goodman A. C., Thibodeau T. G. (2003 Sep). Housing market segmentation and hedonic prediction accuracy. Retrieved from https://www.sciencedirect.com/science/article/pii/S1051137703000317

[8] Gupta R., Kabundi A., Miller S. M. (2009 May). Using Large Data Sets to Forecast House Prices: A Case Study of Twenty U.S. States. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.464.2380&rep=rep1&type=pdf

[9] Khamis A. B., Kamarudin N. K. B. (2014 Dec). Comparative Study On Estimate House Price Using Statistical And Neural Network Model. Retrieved from http://www.ijstr.org/final-print/dec2014/Comparative-Study-On-Estimate-House-Price-Using-Statistical-And-Neural-Network-Model-.pdf

[10] Li Y. (2011 July) Forecasting Housing Prices: Dynamic Factor Model versus LBVAR Model. Retrieved from https://ageconsearch.umn.edu/record/103667/files/AAEA-Forecasting%20Housing%20Prices.pdf

[11] Limsombunchai V., Gan C., Lee M. (2004 March). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. https://researcharchive.lincoln.ac.nz/bitstream/handle/10182/5198/House_%20price_%20prediction.pdf;sequence=1

[12] Morano P., Tajani F., Torre C. M. (2015 Jan). Artificial Intelligence in Property Valuations. An Application of Artificial Neural Networks to Housing Appraisal. Retrieved from https://pdfs.semanticscholar.org/57de/46ad59391c6b120d076863e5387111a60a1d.pdf

[13] Mu J., Wu F., Zhang A. (2014 Aug 4). Housing Value Forecasting Based on Machine Learning Methods. Retrieved from https://www.hindawi.com/journals/aaa/2014/648047/

[14] Muzumdar, P. (2014). Effects Of Zoning On Housing Option Value. Journal of Business & Economics Research (JBER), 9(5), 41. doi: 10.19030/jber.v9i5.9026

[15] Nagaraja C. H., Brown L. D., Zhao L. H. (2011 Apr 14). An autoregressive approach to house price modeling. Retrieved from arXiv:1104.2719

[16] Oxenstierna, J. (2017). Predicting house prices using Ensemble Learning with Cluster Aggregations (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-345157

[17] Park B., Bae J. K. (2015 April 15). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417414007325

[18] Phan T. D. (2018 Dec). Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. Retrieved from https://ieeexplore-ieee-org.prx.library.gatech.edu/stamp/stamp.jsp?tp=&arnumber=8614000