Australasian Transport Research Forum 2021 Proceedings
8-10 December, Brisbane, Australia
Publication website: http://www.atrf.info

Predicting Housing Prices with Ensemble Models

Hao Wu[1], David Levinson[2]

[1]School of Civil Engineering, University of Sydney

[2]School of Civil Engineering, University of Sydney

Email for correspondence: h.wu@sydney.edu.au

1.Introduction

This research models residential housing prices in Sydney, Australia using ensemble hedonic models. The prices of residential properties depend on a wide range of factors, including the functional characteristics as living spaces, and the convenience of transport in reaching urban opportunities. Therefore the ensemble hedonic model is closely related to the transport domain. The combined price of residential property and the cost of transport are subject to constraints of the buyers' budget, creating a trade-off between property price, and transport cost; therefore similar properties are cheaper at more remote locations, because the cost of transport would be higher (Alonso et al., 1964). This trade-off is backed by empirical data (Nelson, 1977). The positive effect of access to employment opportunities on the land value is collaborated by many other research, where distance to CBD (Brigham, 1965), or to highway (Mohring, 1961) is used as proxy for access to jobs. Bus rapid transit (Mulley and Tsai, 2017) and light rail (Mulley et al., 2018) have been identified to have positive effect on Sydney property prices. Both automobile and transit access to jobs contribute to higher property price in Sydney, and transit has a stronger effect than automobile (Rayaprolu and Levinson, 2019), The value for the convenience of transport is reflected in the sales price of residential properties, and accounting for this location-bestowed value is essential in estimating the value of properties.

However, most modeling applications rely on the assumptions of a single model, or compare outcomes from individual models. Ensemble forecasting is a different modeling approach that acknowledges uncertainties in modeling and aims to improve forecast accuracy by combining data and different model outputs. Ensemble models are also capable of presenting model predictions as a range of possible outcomes instead of a singular deterministic number, in order to reflect inherent modeling uncertainties, which makes ensemble models more useful as decision support tools than the single-model approach. Ensemble models have been applied in other fields, most notably in weather forecasting where it significantly improved forecast accuracy (Blum, 2019). There are different types of ensemble models, which are all based on rules specifying how data and models should be combined. These rules vary significantly among different ensemble models, that differ in the ease of implementation, computation cost, and model performance. Some of the rules are used more often than others. Figure 1 shows the range of ensemble models.

Ensemble forecasting intends to extract more information out of available data, and to in-corporate uncertainties in modeling. The resulting ensemble models have higher accuracy, better reliability, and with model outputs that are more useful as decision support tools. The defining characteristic of ensemble models is the combination of outputs from different models, and data from different sources. Philosophically this combination of data and models constitutes an aggregation of information, since different models can extract different pieces of information

embedded within the data (Winkler, 1989); data from different sources also contain non-overlapping pieces of information, that can be combined by ensemble models.

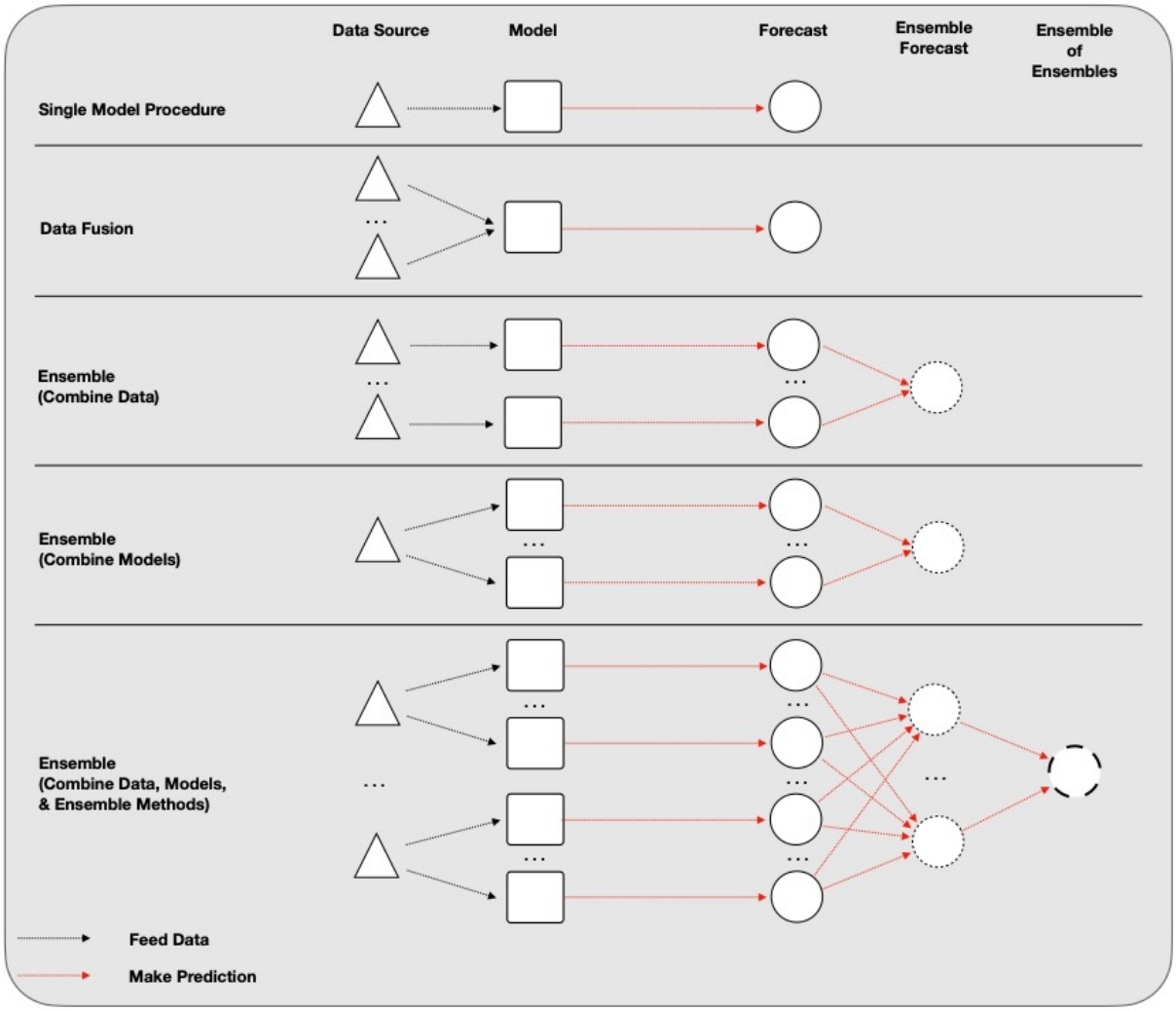

Figure 1. Methods of Combining Data and Models

We model the sales price of houses in Sydney, Australia, using ensemble models (Wu and Levinson, 2021), and variables describing the functional characteristics, and location of the house. Apartments and other residential units are not considered. We compare ensemble model outputs with single model predictions in terms of accuracy, reliability, and usefulness as decision support tools. The objective of this paper is to test different models' ability in extracting information, and to examine the effectiveness of ensemble models in predicting residential property prices.

2. Methodology

We calibrate base and ensemble models to predict Sydney house transaction prices. Each base model consists of a single model, one data input, and produces a single number as model output; both linear and machine learning models are included as base models. Ensemble models combine outputs from base models, and some include multiple data inputs. Models used in this research only consider explanatory variables that are related to the location and functional characteristics of the property; subjective factors such as aesthetics, are left out of the model specifications.
Ensemble models combine base models using predefined rules. We test base models, and three categories of ensemble models in this hedonic application, namely, simple rules, stacking, and ensemble of ensembles.

- **Base models** include five types of models: linear model, classification tree, random forest (RF), gradient boosting machine (GBM), and neural network (NN). The base models independently predict the transaction price of houses, using the same set of explanatory variables and the same training data.
- **Ensemble models with simple rules** combine base model predictions as weighted averages. Two weighting schemes are used, namely the simple averages with equal weights, and weighted average that use model performance metrics (RMSE) of each base model in the training data as weights. Predictions from the same type of base model are assigned the same weight.
- **Meta-learner ensemble models (stacking)** use forecasts from the 5 base models as explanatory variables in predicting the transaction price. The training data is further divided into two portions, one used to calibrate the base models, and the other to calibrate the meta-learners. Three meta-learner ensemble models: linear, RF, and GBM, are trained to combine predictions from base models. A peculiar type of meta-learner uses a RF classifier, and the same set of explanatory variables used by based models, to identify for each residential property, which of the base models is likely the most accurate. For simplicity, the RF classifier is only allowed to choose between two base models: RF and GBM.
- **Ensemble of ensembles** combines different methods of combining base models. Here we use the simple average of the three meta-learner ensemble models for ensemble of ensembles.

The Sydney property transaction data is obtained from the Australian Urban Research Infrastructure Network (AURIN), which records the transaction date, price, location and basic attributes of residential properties. We narrow down the date of transaction to between January 2017 to May 2019, to obtain a sizable dataset, and to limit the effect on housing price from economic fluctuations, and thereby also limiting the analysis to avoid the COVID-19 period. This data records property transactions, which the models are trained on.

Two categories of explanatory variables are used:

- Variables describing the functional attributes of houses.
- Variables measuring location, and the convenience of transport.

The convenience of transport is measured by accessibility to jobs and to urban amenities, and uses actual road network data, and data measuring traffic conditions for driving, and digitized

service schedule for public transport. Percentage of people with foreign origin, walking access to hospitals, and flight noise are also used as explanatory variables. Accessibility is calculated for walking, transit and automobile separately for all 58,819 Mesh Blocks in the Greater Sydney area. Jobs and population data comes from the 2016 census.

3. Findings

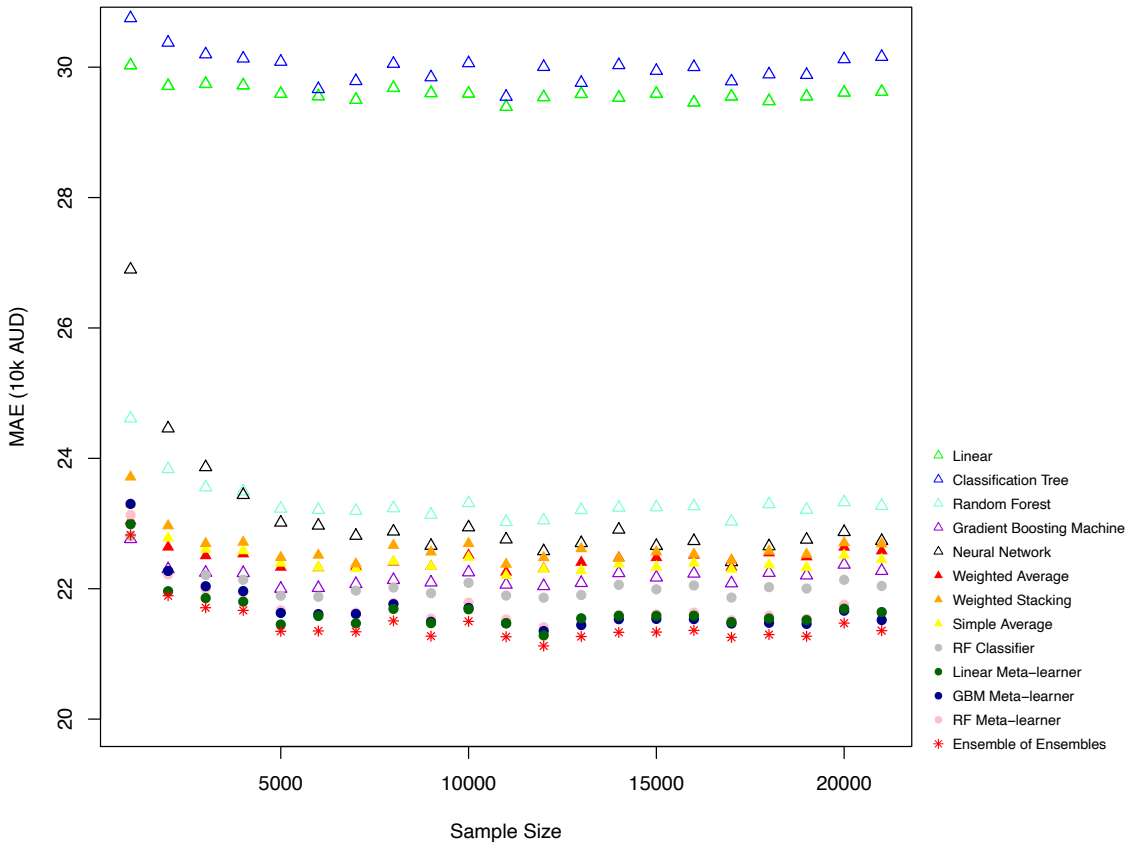

Figure 2. Model performance in predicting house sales price. Mean absolute error of forecast by different models, in testing data. Every dot is the average of 90 experiments, each with 100 testing samples.

We find that ensemble models not only provide more accurate estimates for house price (about 5% improvement) than the single-model approach, but also internalize, and reflect modeling uncertainties as a range of possible model outputs, making ensemble model outputs more useful as decision support tools. In cases with a small training data set available, it may not be possible to calibrate meta-learners, or the inadequately trained meta-learner might have bad performance. Machine learning models generally produce more accurate and more reliable forecasts for the sales prices of houses than linear models (except the classification tree, which is less accurate and not as reliable as the linear regression). Meta-learner ensemble

models are able to improve forecast accuracy and reliability beyond the best base model. Forecast accuracy of ensemble models can be further improved, by combining different methods of combining models (ensemble of ensembles). Given sufficient training data, the ensemble of ensembles is the best ensemble model, and linear combination of based models are generally more robust than other models.

This paper presents empirical evidence for the potential benefit of ensemble models in predicting housing prices. In this research we present the case for using ensemble forecasting to improve transport modeling, which provides a different approach to modeling, and addresses many problems with the single-model doctrine. Ensemble forecasting is more of a paradigm for modeling than a specific modeling method, it enables modelers to view the real-world events (data generation processes) as having multiple possible paths and causes, recognizing some degree of uncertainty.

4. References

Alonso, W. et al. (1964), *Location and land use*, Harvard University Press Cambridge, MA.
Blum, A. (2019), The weather machine: A journey inside the forecast, Ecco.
Brigham, E. F. (1965), 'The determinants of residential land values, *Land Economics* 41(4), 325–334.
Mulley, C et al. Does residential property price benefit from light rail in Sydney? *Research in Transportation Economics*, 67:3–10, 2018.
Mulley, C and C. Tsai. Impact of bus rapid transit on housing price and accessibility changes in sydney: A repeat sales approach. *International Journal of Sustainable Transportation*, 11(1):3–10, 2017.
Mohring, H. (1961), Land values and the measurement of highway benefits , *Journal of Political Economy* 69(3), 236–249.
Nelson, J. P. (1977), Accessibility and the value of time in commuting, *Southern Economic Journal* pp. 1321–1329.
Winkler, R. L. (1989). Combining forecasts: A philosophical basis and some current issues. International Journal of Forecasting 5(4), 605–609.
Wu, H. and D. Levinson (2021) The ensemble approach to forecasting: A review and synthesis.*Transportation Research Part C: Emerging Technologies 132.*
Rayaprolu, H and D. Levinson. What's access worth? a hedonic pricing approach to valuing cities. 2019. (working paper)