Can we predict house prices using known features of each house and a supervised learning approach?

Florence Galliers

18/10/2020

Contents

1 Background:

House prices are an important part of the economy and usually reflect trends in it. They can be influenced by the physical condition of the house and by other attributes such as location (Bin 2004). Prices are important for homeowners, prospective buyers and estate agents, prediction methods could help lead to more informed decisions to each of these stakeholders. Gao et al. (2019) suggest that prediction models may be useful for a few reasons, firstly in narrowing down the range of available houses for prospective buyers. People looking to put their house on the market could use prediction models to look for the optimal time to do so. Prediction accuracy is important.

In most countries there is some form of house price index that measures changes in prices (Lu et al. 2017). This contains a summary of all transactions that take place but not the individual features of each house sold, therefore it cannot be used to make predictions of house price.

Something else to take into account is that it may be difficult for prospective buyers to visualise how square footage measurements of a house are calculated or how this measurement translates into physical size if they have not visited the house themselves. Buyers therefore rely on factors such as the number of bedrooms, bathrooms or house age to get an idea of the value of the house. This analysis will focus on which features of a house have the largest influence on the prediction of house prices. This report does not look at the effect of time on house prices. It is already a well known fact that house prices increase every year (Alfiyatin et al. 2017).

Many house price prediction models have been created using machine learning methods. The hedonic price model is the most extensively researched and uses regression methods (Gao et al. 2019). Machine learning methods use data in a 'training set' to build a model, this model is then used to make predictions on an unseen 'test set' of data. The accuracy of models can be calculated by taking the predicted values from the actual values.

1.1 Objectives:

- Understand which attributes of houses given in the data set can be used to effectively construct a prediction model for house price (dependent variable).
- Minimize the differences between predicted and actual house prices by using model selection to choose the most accurate model.

2 Data:

2.1 Data Description

The dataset chosen for this analysis is from houses sold in 2014 in Washington, USA. It contains the sale price (US dollars) along with attributes of each house such as number of bedrooms, number of bathrooms, etc. There were 4600 observations with 17 variables in the original data set downloaded from [Kaggle] (https://www.kaggle.com/shree1992/housedata). Although the dataset is from 2014, it was a particularly interesting data set because it contains a large amount of information and interesting selection of variables. A more recent data set, or one from the UK could not be found, and so this analysis will go ahead with this data set.

2.2 Data Preparation

The first task was preparing and cleaning the dataset. This served two purposes, firstly to get to know the different variables in the data and existing patterns or correlations between them and secondly to carry out feature selection. Missing values were searched for and removed and any observations in which price equalled zero were removed. The cleaned dataset was exported ready for use in the main analysis. This cleaned data set contained 4522 observations and 12 variables (Table 1).

3 Methods

House price is a continuous variable and so a supervised learning approach was chosen throughout. The dataset was split randomly into two parts, a training set containing 80% of the observations and a test set containing the remaining 20%. The training set will be used to train all of the models and the test set will be used to assess accuracy of the models.

3.1 Linear Regression

To begin, a simple linear regression model was created using price as the dependent variable and square foot living area (sqft_living) as the independent variable. Sqft_living was chosen because it was shown to be the most correlated variable to house price in the exploratory data analysis. A simple linear regression model uses the *lm* function. Linear regression fits a line of best fit which minimises the difference between predicted and actual values.

To see what effect other variables had on this, multiple linear regression was then carried out using all of the other variables available.

It became apparent that sqft_living variable and the multiple linear regression models yielded different RMSE values, suggesting that some of the other variables in addition to sqft_lviing must be contributing to the prediction of price.

3.2 Variable Selection

To look into this further and decide which variables were most influential on price, variable selection methods were explored. Best subset selection, forward selection and backwards selection were all tried. Best subset selection is a method that finds the best combinations of predictors that produce the best fit in terms of squared error. Forward selection is slightly different as it starts with no variables and one by one adds the variable which gives the smallest increase in squared error. Backward selection follows the same idea as forward selection but it starts with a full model, and iteratively removes variables until it leaves a one variable model with the lowest mean squared error.

Using cross-validation in best subset selection, a model with 15 variables was shown have the lowest cross-validation error. These variables were extracted and a linear model created containing only them. This linear model had a lower RMSE than either of the simple or multiple linear models tried above. It was decided that going forward, these variables would be the ones used to trial other methods.

Forward and backward selection yielded almost idential results to each other with only a one variable difference to best-subset selection, so the variables chosen using best-subset selection were the ones used.

3.3 Ridge Regression and LASSO

Ridge Regression and the LASSO are both shrinkage methods.

Ridge Regression utilises L2 regularisation, this adds a penalty to the coefficients that is equal to the square of the coefficients.

LASSO is another coefficient shrinkage technique in which the L0 norm is replaced with the L1 norm. This applies a penalty to the coefficients equal to the absolute value of the coefficients. In LASSO methods, lambda the tuning parameter allows some coefficients to be set equal to 0, in which case they are dropped out of the regression model. In this way it also acts as a selection method for choosing the variables with the most influence. We can change lambda to any value, but in this analysis cross-validation was used to select the best value of lambda.

3.4 Tree Based Methods

Multiple tree based methods were explored in this analysis, starting with one simple decision tree, moving through randomForests, bagging and boosting.

Using just one decision tree gave a RMSE higher than using just ordinary least squares.

3.4.1 Random Forests

RandomForest is a method that combines together multiple decision trees.

3.4.2 Bagging

Bagging stands for bootstrap aggregation. This method uses multiple decision trees (the 'aggregation' part) that are each trained with different data samples, with replacement (the 'bootstrap' part)

3.4.3 Boosting

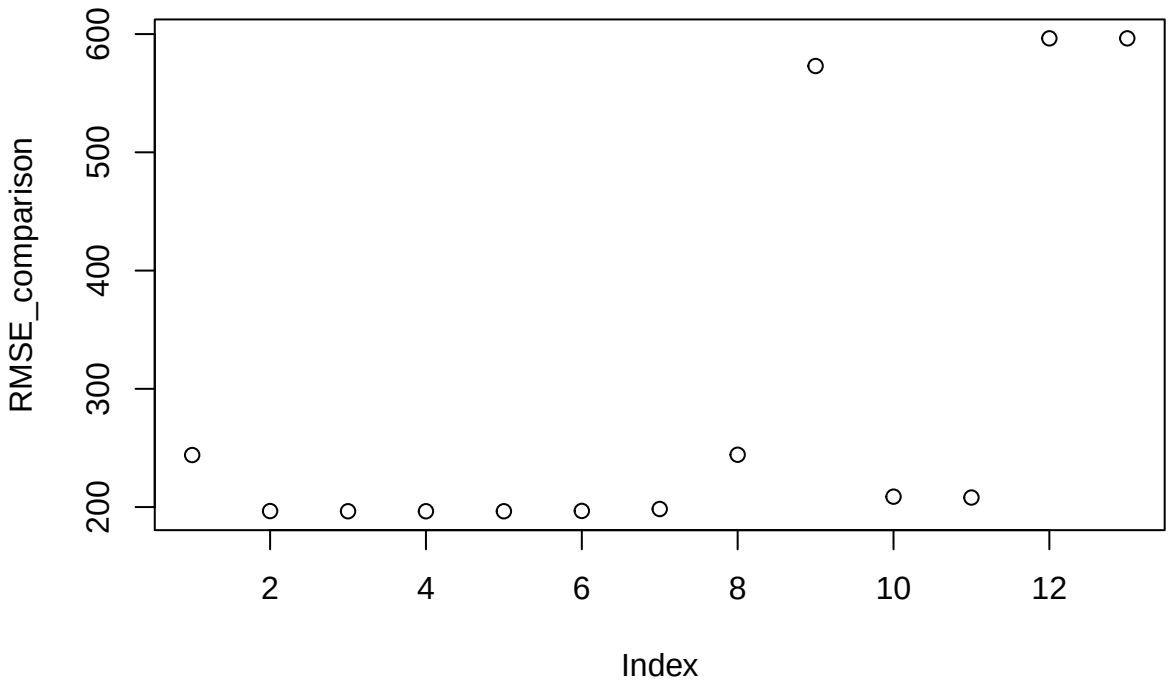
Boosting

4 Results:

```
RMSE_comparison <- c(simple_lm_RMSE,
  multiple_lm_RMSE,
  final_bestsub_RMSE,
```

```
  backward_RMSE,
  forward_RMSE,
  lasso_RMSE,
  ridge_RMSE,
  tree_RMSE,
  rf_RMSE,
  rf_100_RMSE,
  rf_mtry3_RMSE,
  boost_RMSE,
  boost2_RMSE)

plot(RMSE_comparison)
```



4.1 Conclusions:

References

Alfiyatin, Adyan Nur, Ruth Ema Febrita, Hilman Taufiq, and Wayan Firdaus Mahmudy. 2017. "Modeling House Price Prediction Using Regression Analysis and Particle Swarm Optimization." *International Journal of Advanced Computer Science and Applications* 8.

Bin, Okmyung. 2004. "A Prediction Comparison of Housing Sales Prices by Parametric Versus Semi-Parametric Regressions." *Journal of Housing Economics* 13 (1): 68–84.

Gao, Guangliang, Zhifeng Bao, Jie Cao, A Kai Qin, Timos Sellis, Zhiang Wu, and others. 2019. "Location-Centered House Price Prediction: A Multi-Task Learning Approach." *arXiv Preprint arXiv:1901.01774*.

Lu, Sifei, Zengxiang Li, Zheng Qin, Xulei Yang, and Rick Siow Mong Goh. 2017. "A Hybrid Regression Technique for House Prices Prediction." In *2017 Ieee International Conference on Industrial Engineering and Engineering Management (Ieem)*, 319–23. IEEE.