House Prices in New York

CC901E2[a]

[a]The University of Sydney

November 6, 2022

The aim of our investigation is to find the most suitable model to predict house prices in New York. Multiple regression methods, including Backward AIC and Forward AIC, were performed to obtain fitted models. In addition, model stability was examined. Three valid models were found: backward model, forward model, and stable model. In-sample and out-of-sample performances of each model were further evaluated. Analyses highlighted that the stable model produced lower r^2 and adjusted r^2 values, as well as higher RMSE and MAE values. However, it is more stable in comparison to the other two models. Balancing in-sample performances, out-of-sample performances, and stability, we concluded that the stable model may be the most appropriate among the models. With further information about the dataset, a domain knowledge expert may make better judgement on which model to choose.

House prices prediction | Linear regression models

1. Introduction

As defined by the United Nations (UN), "housing is the basis of stability and security for an individual or family" (UN, 2022). Price is an important consideration when purchasing a house. What factors are associated with house prices? What are the best predictors of house prices? To answer these questions, in this report, the data set of house prices in New York was used to find the most suitable model for the prediction of house prices.

2. Data Set Description

The data set was obtained from The Data And Story Library (DASL, 2006), originally sourced from the Mosaic package in R (Corvetti, 2006). It contains 1,734 rows and 16 columns with 16 variables included: 10 numeric and 6 categorical.

These variables provide information about houses in New York, including prices (USD), lot sizes (acres), ages (year), land values (USD), living areas (square feet), percentages of neighbourhood that graduated college (`Pct.College`), whether houses are newly constructed or waterfront (1 stands for true, 0 for false), number of rooms, bedrooms and bathrooms, as well as types of utilities used (such as fuel or heating). However, no information about the "Test" variable was included in the metadata, so it was removed from the data set during the process of analyses.

3. Multiple Regression Models

The scatter plots show that the variables follow linear relationships (Figure A1). Therefore, transformations of the initial data were not made, and the Backward AIC and Forward AIC methods were used to perform multiple regression to find fitted models.

3.1. Backward AIC. For the Backward AIC, the model selection process starts with all variables. It then removes the least informative variables from the fitted model based on how they would affect the AIC value until the lowest overall AIC is reached. The following fitted model was obtained (backward model):

Price ~ Lot.Size + Waterfront1 + Age + Land.Value + New.Construct1 + Central.Air1 + Heat.TypeHot Air + Heat.TypeHot Water + Heat.TypeNone + Living.Area + Bedrooms + Bathrooms + Rooms

3.2. Forward AIC. For the Forward AIC, the model selection starts with no variables other than the dependent variable (i.e. price). It then adds the most informative variables based on the AIC until the fitted model reaches the overall lowest AIC score. The following fitted model was obtained (forward model):

Price ~ Living.Area + Land.Value + Bathrooms + Waterfront1 + New.Construct1 + Heat.TypeHot Air + Heat.TypeHot Water + Heat.TypeNone + Lot.Size + Central.Air1 + Age + Rooms + Bedrooms

4. Model Stability and the Stable Model

Variable inclusion plot and model stability plot are used to examine the stability of the found backward and forward models. During the process, a new stable model was found.

4.1. Variable inclusion plot. The variable inclusion plot (Figure 1) shows that `Land.Value`, `Living.Area`, `Bathrooms`, `New.Construct`, and `Waterfront` are the five most important variables for predicting `Price`. In comparison, `Sewer.Type` and `FuelType` do not provide useful information, as they lie below the path of redundant variables.

The found backward and forward models contain `Land.Value`, `Living.Area`, `Bathrooms`, `New.Construct`, and `Waterfront` parameters, and avoid `Sewer.TypePublic` and `FuelType`.

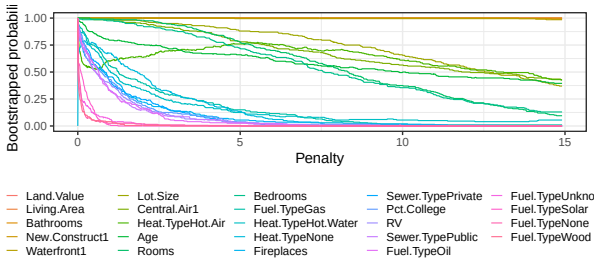

Fig. 1. Variable Inclusion Plot

4.2. Model stability plot. According to the model stability plot (Figure 2), there appears to be a dominant model in each size of **two**, **three**, and **five**. This is demonstrated by one circle being substantially larger than the others in the graph.

The plot further indicates that a model of **size 13** is not stable, due to its small circle size. Indeed, the found backward and forward models are of size 13.

4.3. The stable model. Results in the Variable inclusion plot and Model stability plot sections imply that there may exist a model containing the most important predictors for house prices, and at the same time, of smaller size.

In fact, the following model (stable model) is selected in 78.00% of bootstrap resamples (Figure A2):

Price ~ Waterfront1 + Land.Value + Living.Area + Bathrooms

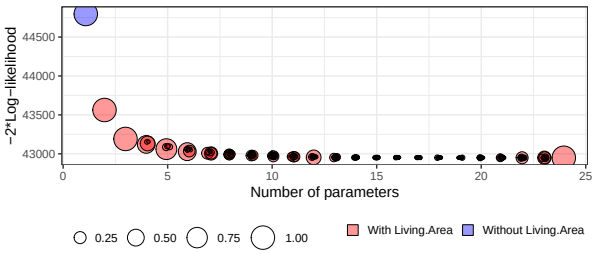

Fig. 2. Model Stability Plot

5. Assumption Checking

Four assumptions were checked for the Backwards AIC, Forwards AIC, and stable models.

5.1. Linearity. The linearity assumption was met. As shown at the beginning of the Multiple Regression Models section, a linear relationship was shown in the scatter plots (Figure A1). Furthermore, there was no obvious pattern shown in the residuals vs. fitted values plot (Figure A3, A5, A7).

5.2. Independence. The independence assumption was likely not met. Properties within the same area are likely not independent, as they will typically have similar attributes due to spatial correlation.

5.3. Homoscedasticity. The homoscedasticity assumption was met. The residuals do not seem to be fanning out over the range of the fitted values in the residuals vs. fitted values plot (Figures A3, A5, A7).

5.4. Normality. The normality assumption can be assumed to be met. While most of the points are reasonably close to the line on the QQ plot, there is some departure from the line at both ends (Figures A4, A6, A8). Due to the sufficiently large sample size, however, we can disregard this and assume normality while relying on the central limit theorem to ensure that the inferences are at least approximately valid.

6. Model Evaluation and Results

In Multiple Regression Models and Model Stability sections, three valid models are found, namely, backward model, forward model, and stable model.

In this section, in-sample performances and out-of-sample performances of the three models were examined.

6.1. In-sample performances. In-sample performances are evaluated by calculating r^2 and adjusted r^2 values of the three models within the data set (Table 1). We found that backward and forward models have the same r^2 (0.655) and adjusted r^2 values (0.652). The stable model have slightly lower r^2 and adjusted r^2 values than the backward and forward models, but very similar (0.633 and 0.632 respectively).

6.2. Out-of-sample performances. The 10-fold cross validation shows that backward and forward models have slightly lower averaged RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) than the stable model (Table 2). Figures of RMSE and MAE further confirm this, despite their large overlaps (Figure A9, A10).

Table 1. r^2 and adjusted r^2

Models	r^2	Adjusted r^2
Backward	0.655	0.652
Forward	0.655	0.652
Stable	0.633	0.632

Table 2. RMSE and MAE

Models	RMSE	MAE
Backward	58190.066	41531.593
Forward	58106.615	41467.353
Stable	59832.246	42709.202

7. Discussion, Conclusion, and Limitations

7.1. Discussion. In-sample performance evaluation indicates that the backward and forward models explain the data slightly more than the stable model, due to their higher r^2 and adjusted r^2 values. However, even though the stable model has lower r^2 and adjusted r^2 values, it still manages to explain approximately 63.27% (according to its r^2 value) or 63.18% (according to its adjusted r^2 value) of the total variation in the house price data, which is reasonable.

Out-of-sample performance evaluation shows that the backward and forward models may be slightly more accurate than the stable model due to their lower RMSE and MAE values. Yet, the differences in the error rates are minor (refer back to Table 2). The accuracy of the stable model is thus ensured.

At the same time, as mentioned in the Model Stability section, the backward and forward models with 13 parameters are not stable. In comparison, the stable model is more stable, as is selected in 78.00% of bootstrap resamples.

7.2. Conclusion. Balancing the results of in-sample performances, out-of-sample performances and stability, we conclude that the stable model may be the most suitable among the three models.

7.3. Limitations. The chosen stable model is a compromise between stability and accuracy. In certain contexts, however, it might be preferable to choose accuracy over stability. Additionally, as mentioned in the Assumption Checking section, the independence assumption may not be met. Therefore, if more information about the data set is provided, a domain knowledge expert may make better judgement on which model to choose.

References

Corvetti C (2006). *Houses in Saratoga County*. R package version 1.17, URL https://www.saratogacountyny.gov/departments/real-property-tax-service-agency/.

DASL (2006). *Housing-prices-GE19*. Accessed November 06, 2022, URL https://dasl.datadescription.com/datafile/housing-prices-ge19.

UN (2022). *Housing is a right, not a commodity*. Accessed November 06, 2022, URL https://www.ohchr.org/en/special-procedures/sr-housing/human-right-adequate-housing#:~:text=Housing%20is%20the%20basis%20of,in%20peace%2C%20security%20and%20dignity.