The University Of British Columbia
Department of Statistics

STAT 406 Project Report

Modelling and Predicting the Greater Vancouver's Housing Price

Authors:

Yiwen Jin
HaoYang Xu
Shukai Yan
Jiayan Yang

Student Number:

18147926
33632134
43533141
16237166

Professor:

Matias Salibian

Nov.30th,2017
Academic Year 2017/2018

1 Introduction

The Greater Vancouver's real estate market has been torrid recently with the local government trying to cool off the market. We wish to find relevant factors affecting the price and use the data to train a model for prediction. This project can provide investors with a quantitative picture of the real estate market. For the rest of this report, we will summarize details of all interesting explanatory variables, demonstrate methods to predict house prices, and analyze all models' outputs.

2 Brief Summary

We aim to find factors and refine them to gather our explanatory variables. We will construct the optimal learner to predict the average housing price for the month. This project intends to establish a model, which gives a best prediction, and might be helpful for people to understand more about our real estate market, and avoid housing market bubbles.

3 Project Details

3.1 Data Description

We have 16 sets of monthly data for Great Vancouver from January, 2009 to October, 2017. Our 16 explanatory variables are related to Vancouver monthly economy measurements (e.g. CPI, exchange rate), monthly real estate (e.g.rent, sales, listing numbers), and environmental conditions. The response variable is the average price of detached houses in Greater Vancouver. In total we have 105 observations. All variables are recorded in the Greater Vancouver area. Full data descriptions are in Table 1.1. We also scaled the variables by 1000s for an easier interpretation of the data and model results.

3.2 Methodology

The models we are going to build are linear regression, Elastic Net, random forest, and Bagging. The basic idea of this project is to compare the mean squared prediction errors. We will run a 5-fold cross validation 50 times to predict the MSPE. Taking the average of of the 50 MSPE will avoid the random extreme cases. We will build different models, and use these models to predict the response variable, the average Greater Vancouver's detahced house sale price. By comparing the mean squared prediction error, we choose the best prediction model with the smallest mean squared prediction error. More details will be provided in the Analysis section.

4 Analysis

4.1 Linear Regression

Linear regression is the simplest model in this project which gives a linear function of explanatory variables. For linear regression, we will use *stepAIC* in R to select features. Not like residual of sum squared and R-squared, which always decreases when we add features or make the model more complex, AIC of the model is about the likelihood of parameters, and is constrained by the number of parameters we are using. Therefore, we can select the most important features by choosing the model with the smallest AIC. and forward stepwise selection gives the smallest MSPE. After checking the residual plot, the model seems to be reasonable. The boxplot of the 5-fold cross validations is shown in the Section of Appendix-Tables and Figure. We obtained a mean of MSPE of 61.50 for stepwise model, 55.49592 for full model.

4.2 Elastic Net

Elastic Net is a combination of Ridge Regression and Lasso Regression. It handles cases where the explanatory variables are highly correlated to each other. It can take both advantages. Ridge regression uses all of the explanatory variables, but gives some variables a coefficient close to zero, and Lasso regression builds a sparse model with less explanatory variables. We use function

cv.glmnet which takes parameter of alpha. If alpha equals to one, the function builds a best Lasso regression based on cross validation, and if alpha equals zero, it is a ridge regression. We set a list of 11 possible values of alpha including zero and one (e.g. 0, 0.1, 0.2, ... 1). Then, compute these 11 Elastic Net models. We choose CV to choose the optimal alpha (0.8) with the smallest MPSE of 52.81. Finally, we use the optimal alpha to produce 50 runs of MSPE. We obtained a mean MSPE of 57.21 for elastic net model, 57.32 for Lasso model and 55.71 for ridge model.

4.3 Random Forest

Random forest algorithm has a strong resistance to overfitting and it uses only parts of the features. Random forests build a large collection of de-correlated trees which is similar to bagging algorithm. We create random forest model with different trees and compare them by using different number of trees in the model as shown in Figure 3 on page 4. The smallest MSPE is 56.6 and the best number of trees is 500.

4.4 Bagging

Pruning a single regression tree for prediction is very unstable and sensitive results in a way that every re-run of the regression model could yield a different result which can be seen from Figure 2 on page 3. To deal with this, we introduce the bagging tree method to fit the data. For a pruned single tree, we obtained MSPE 77.45246. For the bagging trees, we obtained MSPE 53.02581. Bagging trees gave better results.

5 Conclusion

By comparing all the models we built, we obtain all the mean square prediction error (MSPE) values. To obtain the optimal model, we will choose the best model of prediction with the smallest averaged MSPE. Based on the above results, the optimal model for our data is Bagging Tree Regression with MSPE 53.02581. Considering the boxplot results, we can conclude that the bagging trees model has the smallest variance which turn out to be the most stable results.

For improvement, we can consider other criteria for measuring the prediction error such as a Mean Absolute Error and Normalized Mean Squared Error.

6 References

Real Estate Board. www.rebgv.org. Accessed 18 Oct. 2017.
Bank of Canada. www.bankofcanada.ca/. Accessed 18. Oct. 2017.
Government of Canada. www.canada.ca/en.html. Accessed 18 Oct. 2017.
Regional Data and Statistics. www.metrovancouver.org. Accessed 18 Oct. 2017.
Province of British Columbia. www2.gov.bc.ca/. Accessed 18 Oct. 2017.
ICBC Statistics. www.icbc.com/about-icbc/newsroom/Pages/Statistics.aspx. Accessed 27 Nov.2017.
Translink.www.translink.ca/About-Us/Media/2017/July/Record-Ridership.aspx. Accessed 27 Nov.2017

7 Appendix-Tables and Figure

Figure 1: Model Comparison

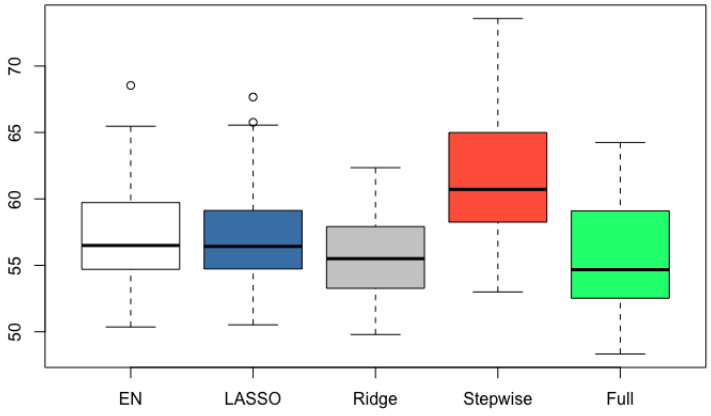

Figure 2: Tree - Bagging tree MSPE

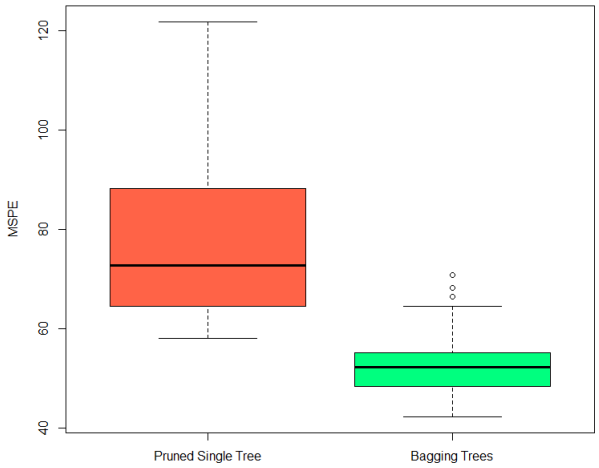

Figure 3: Random Forest

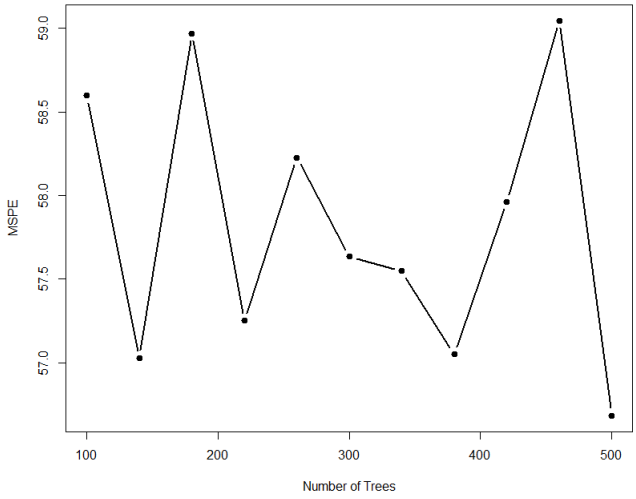

Table 1: Data Description

Variables Name	Descriptions	Unit
AvgSold Price	the average price of detached houses in Great Vancouver area.	CAN Dollars
Licensed Vehicles	The number of vehicles has been activated by valid licensed drivers.	NA
Total Labour Force	Number of individuals in an economy who either are employed or are seeking employment.	NA
Unemp -loyment Rate	The ratio of number of individuals in an economy who are unemployed and the number of total labour force.	NA
CPI	A measure of changes in the price level of market basket of consumer goods and services purchased by households.	NA
Retail Sales	Retail sales are purchased of finished goods and services by consumers and businesses.	1000 CAN Dollars
Max Temperature	The highest weather temperature recorded.	degrees Celsius
Precipitation Accumulation	The amount of liquid(rain,snow) occurred.	mm
Crimes	The number of crimes(personal,properties,etc.) occured in Great Vancouver.	NA
Immigration Population	How many population coming to live permanently in Great Vancouver each month from a foreign country.	NA
Exchange Rate	How Canadian dollars valued compared to other currencies.	NA
Mortgage Rate	the rate of interest charged by a mortgage lender.	NA
Total Transit Ridership	The passengers who use a given public transportation system, as buses or skytrains, or the number of such passengers.	NA
#ofsales	The number of houses sold each month.	NA
New Listings	The number of new house being constructed.	NA
Season	The season (Spring,Summer,Fall,and Winter)	Categories