PREDICTING CONDOMINIUM PRICES IN GREATER VANCOUVER

Prepared by:

Abdullah Farouk,
Mengling Zhou,
Weining Hu,
Xin (Kathy) Zhao

Summary

The market for condominiums (condos) in Greater Vancouver continues to boom over time. The interaction between multiple factors makes it difficult to determine the condo's final selling price. In this report, we propose the use of a hybrid Random Forest model to predict the price of a condo. A novel feature of our model is its use of predictions from a Vector Autoregression Model (VAR) to capture time trends in average condo prices. We then provide a brief guide on how to use our BC Condo Shiny application. This app allows users to use our model to make predictions of prices of condos as well as showing dynamic visualizations. Finally, we conclude with a comparison of the performance of our model with that of the model built by Stat 450 students. We observe a lower prediction error when using our model to predict condo prices in 2018 relative to theirs.

Introduction

Greater Vancouver has one of the hottest condominium markets in the world, with average condo prices soaring through time, especially since the year of 2015. Condo prices in Greater Vancouver are affected by a multitude of factors, such as floor area, bylaw restrictions, the distance to the nearest Skytrain station, and even the currency exchange rate. These factors interact with the condo prices in a complex manner. Thus, investors and home buyers are interested in estimating condo prices in order to make a reasonable offer and have an idea of the most influential factors on a condo's price. Considering our clients' interest, we have built an easy-to-use condo price prediction shiny app along with this report, which we have written as a user manual to it and as a brief explanation of our model. Due to the clients' needs, our predictions focus on three subregions: Collingwood, Metrotown, and Whalley. Data used to train the models are obtained from realtors' MLS website, containing all the sales record ranging from January 2005 to March 2018.

The upcoming sections proceed as follows. First, we describe how we built our prediction model. This is then followed by instructions on how to use our BC Condo shiny application. Finally, we conclude with a comparison of our results with those obtained by Stat 450 students who are also a part of this project.

Proposed Statistical Analysis

Hybrid Random Forest Model

To predict condo prices in Vancouver we have built a hybrid Random Forest model that incorporates predictions from a VAR model. VAR is a stochastic process model that captures linear interdependencies among multiple time series. These models allow for the predictions of several time series variables simultaneously. We chose this approach for the following reasons.

- Random forests are well suited to making predictions in situations where the response variable (price) may depend nonlinearly on interactions between various variables of interest.

- Condo prices have both a yearly trend and seasonal trend. Thus, in order to predict condo prices accurately, it is important to account for trends in average prices over time.

Vector Autoregression Model (VAR)

We use a VAR model to overcome initial difficulties we faced when trying to use an ARIMA model to capture the effects of foreign currencies (e.g. CAD vs USD) and government policies on condo prices.

This is because data on foreign currencies form a time series themselves whose values cannot be known in advance. Thus, to predict the average price of a condo in December 2018, we have to first predict the values of our foreign currency variables for December 2018 which increases our prediction errors.

We have built a VAR model with variables for average condo prices and foreign currencies as our three evolving variables and government policy as our other independent variable.

The diagram below shows how our VAR model predicts house prices in Metrotown. The points in black represent observed prices whilst those in red are our predicted prices. The shaded dark grey region on the graph represents confidence intervals over our predictions (i.e. we expect predicted prices to be in that interval with 90% probability). We can see that most of the observed prices lie in our prediction interval of real prices.

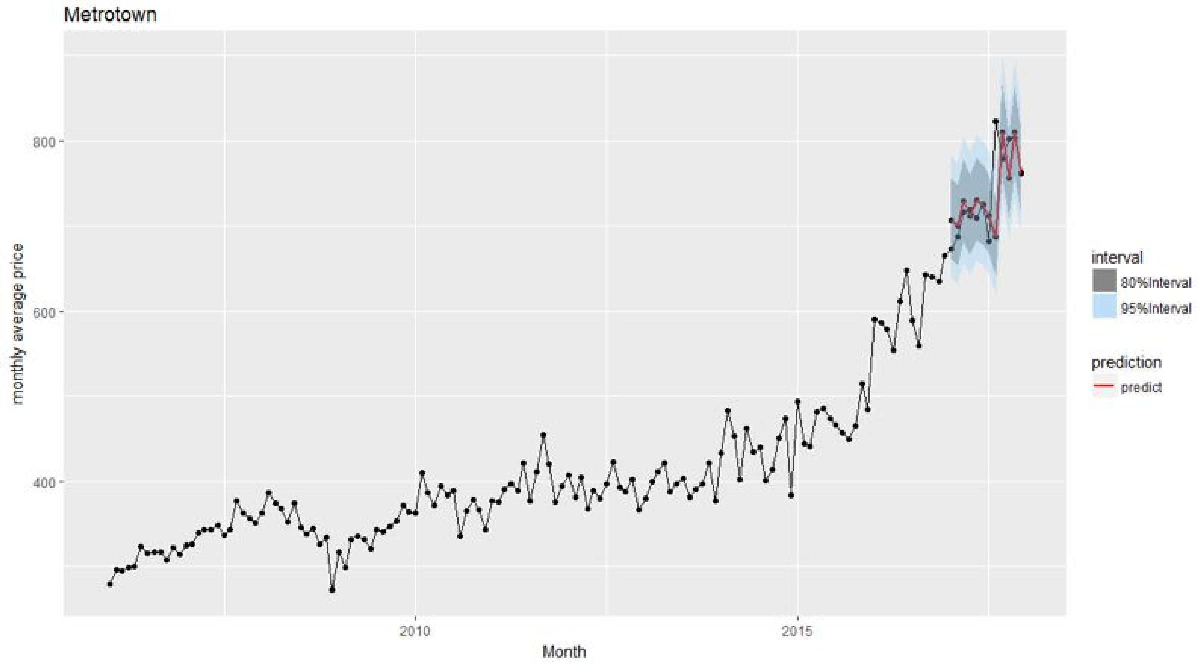


Figure 1. Time-series analysis of Metrotown

BC Condo Shiny App

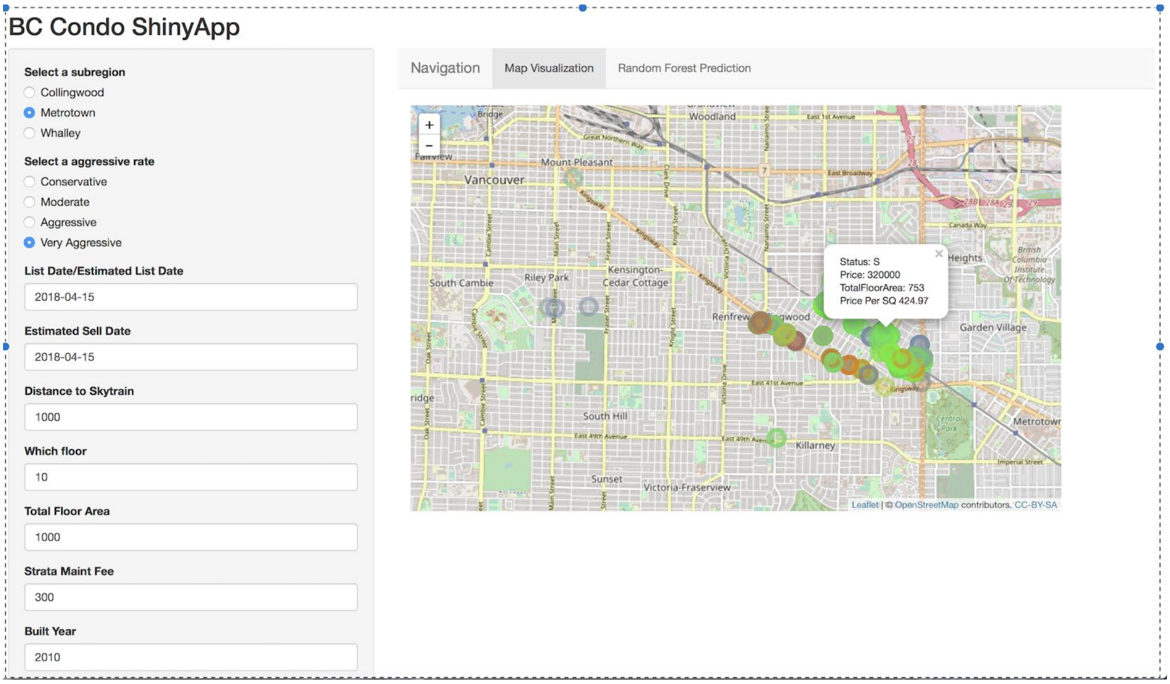
In this section, we demonstrate how to use our BC Condo Shiny app, which allows users to:

- visualize condo prices across three different neighborhoods in Vancouver,
- make predictions of house prices by specifying values of the different inputs in our model themselves.

Our Shiny App has two parts.

1. Heat Map Visualization

In this section, the algorithm behind the app parses the addresses of condos from the dataset provided and uses a Google Map API to get the latitude and longitude for these condos. By using these latitudes and longitudes, we add markers onto a dynamic map in real time. When users hover over these markers, basic information such as Status, Price, Total Floor Area as well as Price per square feet is shown for that specific condo. In addition, the colors of these markers change with prices, i.e. the higher the price of a condo the warmer (darker) the color of the marker.



BC Condo ShinyApp

Select a subregion
○ Collingwood
● Metrotown
○ Whalley

Select a aggressive rate
○ Conservative
○ Moderate
○ Aggressive
● Very Aggressive

List Date/Estimated List Date: 2018-04-15

Estimated Sell Date: 2018-04-15

Distance to Skytrain: 1000

Which floor: 10

Total Floor Area: 1000

Strata Maint Fee: 300

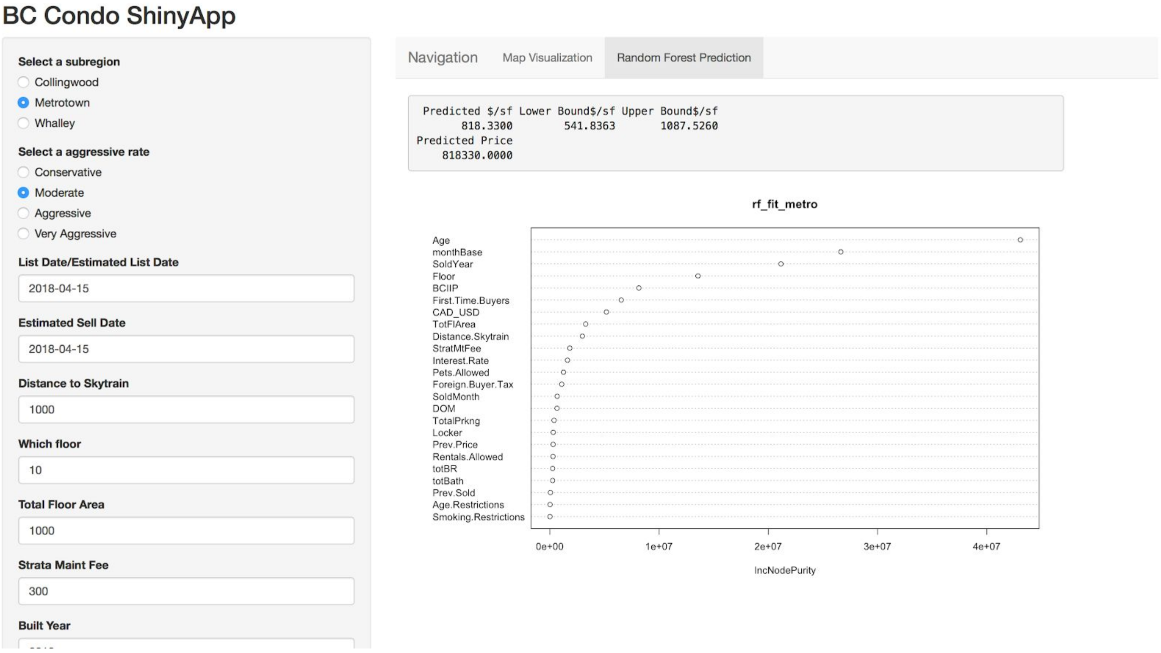
Built Year: 2010

Status: S
Price: 320000
TotalFloorArea: 753
Price Per SQ 424.97

Figure 2. Heat Map Visualization

2. Random Forest Prediction

In this tab, users can type in values for different input arguments on the left-hand side of the screen. Our random forest model uses those input values to give a prediction for the price per square feet, the overall price, as well as the 95% prediction intervals (lower bound, upper bound) for the price/sf Additionally, a corresponding Important Factors plot is displayed to make it easy for users to identify factors with the largest influence on prices.



Select a subregion
○ Collingwood
● Metrotown
○ Whalley

Select a aggressive rate
○ Conservative
● Moderate
○ Aggressive
○ Very Aggressive

List Date/Estimated List Date
2018-04-15

Estimated Sell Date
2018-04-15

Distance to Skytrain
1000

Which floor
10

Total Floor Area
1000

Strata Maint Fee
300

Built Year

```
 Predicted $/sf Lower Bound$/sf Upper Bound$/sf
       818.3300        541.8363       1087.5260
Predicted Price
    818330.0000
```

Figure 3. Random Forest Prediction

There are 24 boxes for users to either type in or select values to be fed into the model. It is initialized using default values. Users can then update these values either by manual specifications or by selecting one of the many options listed. As the model is trained on previous records, extreme inputs by a user may lead to insensible predictions.

A novel feature we have included in our random forest model is letting users decide how aggressive or conservative the model's predictions should be. Given the dynamic nature of Vancouver's real estate market, this feature allows users who are bullish and have an aggressive buying strategy to get a price point that matches their risk appetite.

Here we briefly address some of the limitations of our model:

1. Our predictions are only reliable for condos with an estimated sell date on or before Dec 31, 2018. Predictions after this time are not available.

2. Predictions are based on government policies (i.e. BCIIP, first-time buyer, foreign buyer tax) currently in effect. Any amendments to existing policies may invalidate our predictions (e.g. cancellation of the foreign buyers' tax policy). However, this can be easily adjusted by revising the background server.R file.

3. Users may not see any change in predicted prices when changing the values of certain input arguments. This is because these factors are less important in the prediction process and therefore large changes in their values may not affect the model's predictions. Users should not be alarmed by this as it is not an indication of a failure in the model.

4. Users may find some nonsensical predictions when adjusting inputs in the app. For example, holding all other variables constant, increasing the number of bedrooms may lead to a decrease in the

predicted price. We suspect the error may come from three parts: (1). insufficient data points for those set of inputs; (2). errors in monthly average predictions from our VAR model; (3). other potential useful variables are not considered in the model.

Results

In this part of our report, we compare our predictions with those obtained by Stat 450 students. We first compare predictions from our VAR model with those from the ARIMA model used by them. We then compare predictions from our hybrid Random Forest model, which incorporates VAR predictions, with those from the Random Forest model used by them.

To perform comparisons of the VAR and Hybrid Random Forest Models with those developed by Stat 450 students we use two metrics as measures of prediction errors.

- Mean absolute percentage error (MAPE). It measures the absolute difference between our observed and predicted values, divides this difference by the absolute values of our observations and then expresses this value as a percentage out of 100.

- Mean square prediction error (MSPE). It measures the expected squared distance between our true value and our predictions for that specific value. We only use this metric in the evaluation of our random forest model.

1) VAR Model

To make predictions using VAR we incorporate a rolling window, which uses all the information available. That is, we use the first observed *t* data points to predict the *(t+1)*th data point. We then use the observed *(t+1)th* data points to predict (*t+2)*th and so on. To test our model's predictive capabilities against that of the univariate ARIMA model, we use all data points from 2017 as our test set and evaluate its prediction errors using MAPE. The results from this comparison are displayed in the table below.

Table 1. Comparison of predicted prices (CAD$/sqft) between ARIMA (STAT450) and VAR (STAT550) using the data before 2017 to predict the monthly average prices of 2017.

	ARIMA (STA450)			VAR (STAT550)		
Date	Collingwood	Metrotown	Whalley	Collingwood	Metrotown	Whalley
Jan-17	599.60	662.43	336.47	597.58	707.15	354.48
Feb-17	601.74	666.32	337.73	630.47	699.96	366.68
Mar-17	603.93	671.88	338.99	649.07	728.58	409.58
Apr-17	606.13	674.49	340.24	665.45	711.65	418.50
May-17	608.32	676.21	341.50	657.76	730.21	401.63
Jun-17	610.51	679.14	342.76	666.19	723.65	445.57

Jul-17	612.71	682.53	344.01	696.07	712.32	441.86
Aug-17	614.90	685.42	345.27	709.11	686.66	447.61
Sep-17	617.09	688.09	346.53	727.79	809.71	465.90
Oct-17	619.28	690.96	347.79	733.67	757.10	503.41
Nov-17	621.48	693.93	349.04	789.95	809.73	502.83
Dec-17	623.67	696.83	350.30	775.01	761.52	492.93
MAPE	0.1205	0.0772	0.2479	0.0237	0.0371	0.0563

We observe that the MAPE of our VAR model is much smaller than that of the univariate ARIMA model. This suggests that our method can be used to improve the accuracy of condo price predictions.

2) Hybrid Random Forest Model

To compare prediction powers between the hybrid random forest model and the classical random forest model, we simulate two different scenarios below. In this comparison, we do not perform cross-validation to estimate the prediction error because randomly dividing the data may lead to future trends in prices being seen and therefore used by our model to predict future condo prices. Table 2 and Table 3 below detail results from comparisons under each of the two scenarios. Additionally, we include a 95% prediction interval (PI) coverage rate for each subregion to compare within the model.

Scenario 1: *Using the data before Sept 2017 to predict the last three months in 2017*

Table 2. Comparison of predicted prices (CAD$/sqft) between hybrid random forest (STAT550) and classical random forest (STAT450) using the data before Sept 2017 to predict the monthly average prices of the last three months in 2017.

		MSPE		MAPE		95% PI Coverage Rate
		Value	% Reduction	Value	% Reduction	
Collingwood	STAT550	912.21	40.48%	0.016	53.31%	95.03%
	STAT450	1,532.66		0.035		
Metrotown	STAT550	768.87	41.16%	0.017	49.23%	95.19%
	STAT450	1,306.80		0.034		
Whalley	STAT550	1,171.88	19.38%	0.029	35.76%	92.46%
	STAT450	1,453.65		0.046		

Scenario 2: *Using the data before Jan 2018 to predict the first three months in 2018*

Table 3. Comparison of predicted prices (CAD$/sqft) between hybrid random forest (STAT550) and classical random forest (STAT450) using the data before 2018 to predict the monthly average prices of the first three months in 2018.

		MSPE		MAPE		95% PI Coverage Rate
		Value	% Reduction	Value	% Reduction	
Collingwood	STAT550	12,912.41	39.70%	0.101	27.74%	79.07%
	STAT450	21,412.56		0.140		
Metrotown	STAT550	7,194.61	42.72%	0.079	27.24%	88.00%
	STAT450	12,560.98		0.108		
Whalley	STAT550	9,953.68	43.74%	0.148	28.96%	78.21%
	STAT450	17,693.13		0.208		

Both scenarios indicate incorporating VAR predictions of monthly average prices in a classical random forest model could achieve large improvements in prediction power. However, we notice that our model performs worse in Scenario 2 compared to Scenario 1. We suspect this is due to errors in VAR predictions, used by the model, over this time frame.

Conclusion

In this report, we describe models we built to predict condo prices in three subregions in Greater Vancouver. They offer improved predictions in comparison with the models built by students in 450, evidenced by lower prediction errors relative to those obtained from the models used by the students in 450. We have also included instructions on how to use our BC Condo Shiny app. This app allows users to utilize the model to predict house prices for condos with specific features.