Vancouver Housing Market Decoder

Yuyi Zhou, Junbo Bao, and Yabin Guo

1 Motivation and Background

Vancouver, known as one of the most beautiful cities in the world, attracts millions of people to move in every year. As a result, Vancouver's housing market is always very hot. Buying or Selling a property is not only for living, but also for investment. In the housing market, people play different roles. They can be buyers, sellers, realtors and mortgage managers. A buyer is interested in buying a property with a reasonable price. A seller is interested in how much his property is worth. The change of property value can affect everybody. People want to fully understand the property value and future moving trend so that they can make inform decisions. According to this projects mentor Marco Wu, currently there is no such tool in the market that using machine learning to predict property price and future trend. Our project is the first one applying machine learning models to achieve that.

2 Problem Statement

In order to help people fully understand vancouver housing market and make inform decisions, the problem states from two perspectives. First we need to predict a propertys current value and predict future moving trend. Besides, we need to create a interactive UI for users to easily access the predicted results. Based on these two perspectives, we created three user scenarios.

- The user scenario for seller
  We need to predict the listing price so that the seller knows how much his property is worth and receives a recommended listing price if he wants to sell.
- The user scenario to predict future market trends
  All type of users are interested in the future market trends. We need to predict the future trend for different type of properties in different areas so that users can see the values of which area of properties are increasing or decreasing.
- The user scenario for buyer
  We need to predict the purchase price so that the buyer knows how much he needs to pay to get the property he wants.

Solving problems for these three scenarios are very challenging. Predicting a property value and future change is a hard topic, especially with high accuracy. The accuracy is also highly related to the type and amount of data. However, we didnt have initial dataset. The data we can collect from public are rew listing price, bc assessment value for the properties in city of Vancouver and monthly benchmark property price. We are missing real-time property sold price which is a very important factor that can affect the accuracy of the prediction. We did a lot of effort to work around the lacking data problem and tried to build a high accuracy model based on the data we can collect.

Data pre-processing is also a challenging part. We spend almost 80% time in data preprocessing plus data collection. The data retrieved from online are from different sources and are raw data. We need to integrate them together, remove outliers, add missing value, do feature engineering and convert data into the format that can be fit into models.

In terms of model evaluation, we did a lot of research on how to improve our models accuracy and tried different models since the model accuracy is the most important fact in our project.

In addition, UI is another challenging part. We aimed to create a UI that can be easily accessed by users. The users can enjoy using our tool to retrieve the information they are interested, estimated listing price, estimated purchasing price and property value future moving trend.

- *Y. Zhou, J. Bao, and Y. Guo are with Simon Fraser University.*

3 Data Science Pipeline

The whole data science pipeline is shown in Fig. 1. We will discuss each component in details in this section.

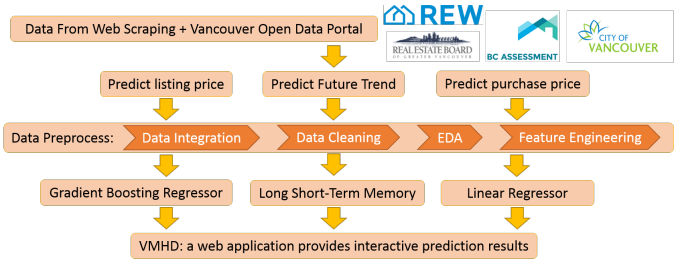


Fig. 1: Pipeline

3.1 Data Collection

Initially, we dont have any existing dataset that can be used directly. Hence we have to start from crawling online data. Based on the user scenarios discussed in the previous section, we decided to collect the BC Assessment data from bcassessment.ca, 2018 Vancouver tax report from City of

Vancouver open data, property listing information from rew.ca, and benchmark prices from real estate board of great Vancouver. Data from each source has a different group of parameters of the properties. It is necessary to collect all data first and try to combine them afterwards.

3.1.1 BC Assessment data

BC assessment (https://www.bcassessment.ca/) is the only website possessing the purchase price we can find. To avoid missing any valuable information, we decided to store the whole web page onto the disk. The main challenge we came across in this part was the Captcha examination of the website. To overcome this problem, we used python selenium package to disguise the crawling program as a web browser. In order to avoid requesting the website too frequently, we made the program sleep for a while and restarted the browser occasionally. The data we scraped from BC assessment has about 200k records. All the Vancouver housing data was successfully collected since we used the property ID (PID) as the searching criteria. The PID of Vancouver properties can be completely retrieved from the 2018 Vancouver property tax report. Among the collected data, 14.5k properties have transaction history within the past three years. The whole scraping process lasts for days even we have tried to crawl the data as fast as possible. The reason lies in the fact that Captcha would make our program stuck if the crawling speed increases even a little bit.

3.1.2 Rew property listings

We collected all the listing information of properties in greater Vancouver from https://www.rew.ca/ The listed properties change everyday. As for our project, we want to obtain as more listings as possible and see if there is any property reposted with a different price. Consequently, we decided to crawl the data twice with an interval of one month. The total number of distinct listings we collected was 12906 after removing the identical records.

3.1.3 Vancouver tax report

The Vancouver tax report is available from 2006 to 2018 in excel format. We downloaded them directly from http://data.vancouver.ca/datacatalogue/propertyTax.htm

3.1.4 Property Benchmark Prices

The real estate board of Vancouver (https://www.rebgv.org/mls-home-price-index) provides property benchmark values from 2015 Jan to 2018 Feb, by month. A benchmark price is the estimated sale price of a benchmark property. Benchmarks represent a typical property within each market. We scrapped data for Vancouver East and Vancouver West. This data is used to predict future trends.

3.2 Data Cleaning, Data Extraction and Data Integration

3.2.1 BC assessment data

The raw data we got from web scraping was in html format. Python BeautifulSoup package was applied to extract all the features and convert them into tabular format. Features from 16 fields can be extracted. To combine the BC assessment scraped data with the Vancouver tax report, we joined two tables with PID which is the unique identification of a property. Different year of tax report was used to join depending on the year the property was sold. For instance, the record of a property sold in 2014 will be joined with the 2015 tax report. Note that the tax report reveals the property tax information of the previous year. Missing value was dealt with accordingly. For most features, filling the field with zero is in accordance with the reality. The combined data contains 47 features. We selected 19 features for next step of analysis. Some features contain values in an inconsistent format. For example, the land size of some properties were represented as the format of 32.99 x 122.04 Ft, and some properties were represented as 3986 Sq Ft. We converted them into the same format, an integer representing the square feet in our case. Additionally, we manually created some features. For instance, the square feet of a house was calculated by summing up the area of basement finish area, first floor, and the second floor. The first three letters of the postal code were used to represent the subarea. We also extracted the house type from the description feature of our data. After the preprocessing steps above, we acquired 17 numerical features and two categorical features of each properties.

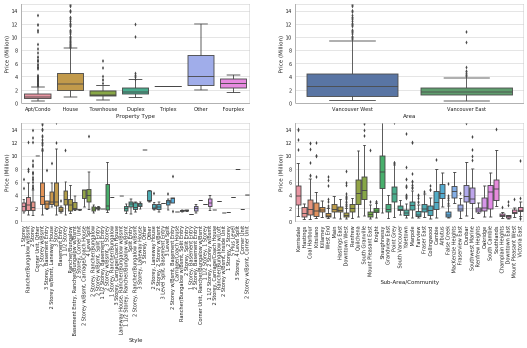
3.2.2 rew data

The raw data scraped from rew.ca was also in html format. We used the similar procedure to convert the raw data into tabular format. One shortcoming of rew data is that it doesnt contain the tax information of the property. Unlike BC assessment data, there is no such a feature like PID in rew data for us to join the tables. To integrate this information into the table, we combined rew data and Vancouver tax report with simplified entity resolution approach. We extracted the street numbers from both tables. In our criteria, the street number of two property must be identical. After that, the Jaccard similarity between the addresses from two sources was calculated. A threshold was used to filter the mismatched records out. In this manner, we found the corresponding records in tax report for most of the records of rew dataset (more than 90 percents). As for the missing values, we made different treatments. For instance, the strata fee of land-type properties were filled with zero simply because land-type properties dont have strata management company. Two dimensions of the lot size were estimated by calculating the root of the lot size. Eventually, 2.5k records were obtained. We didnt consider further combining this combined data with the BC assessment data for the reason that a property being listed in rew.ca can hardly had a transaction within the past three years.

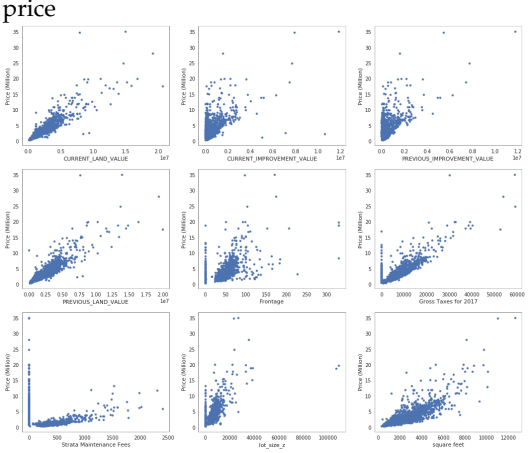
3.3 Data Analysis and Model Training

3.3.1 Seller Scenario

For seller scenario, our goal is to predict the listing price for the seller when she/he wants to sell her/his property. The dataset we used to train the model was the integrated rew dataset. The data of this dataset comes from the actual listing price of the properties in the market. This makes sense since a seller cares more about the listing prices of similar properties than the assessed value of her/his property. To extract the meaningful features, correlations between each



(a) Each category is correlated with the listing price



(b) Most of numerical features show strong positive linear correlation with the listing price

Fig. 2: Investigation of correlation between each feature and the listing price

numerical features and listing price were plotted. Clearly, all nine features shown in Fig. 2a have strong positive relationship with listing price. In addition, each categorical variable was also plotted against the listing price (Fig. 2b). Finally, we picked twelve numerical features and six categorical features to build the predictor. The categorical feature was transformed into the vector with one hot encoding method. Eventually, approximately 2.5k records collected within one month were used to build the regressor. We tried several regressors from python sklearn package, including linear regression, neural networks, and gradient boosting regression. For each model, we preserved 10% of the whole data as testset. We performed 5-fold cross validation on the remaining 90% of data to tune the hyperparameters. The gradient boosting regressor [1] achieved the best R^2 score on the testset ($\sim$0.964). The selected model was saved onto the disk for future predictions.

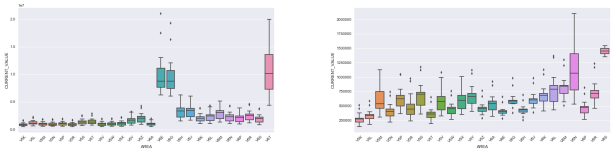
3.3.2 *Predict Future trends Scenario*

There are two parts in predicting future trends. One part is that we used benchmark prices from 2005 Jan to 2018 Feb(158 data points in total) to predict the benchmark price from Feb 2018 to May 2018 so that we can get an overview trend of different type of properties in Vancouver East and Vancouver West. The other part is to combine tax report from 2006 to 2018 to predict average bc assessment value for land and strata properties in different areas in 2019. The area is divided by the first three digital letter of a post code.

In the first part, its a typical time series problem and we applied LSTM[2] model from Keras python library to make prediction. There were a few data preparation steps to transform the data into some formats to fit into the model. After training and getting the results, we need to recover the data back to get the expected result. To predict future trend, the data needs to be stationary which means any increasing trends need to be removed. We transformed data to stationary by getting the differences between the current time point value and the previous time point value. The LSTM input has to be between -1 and 1. So we normalized the data using MinMaxScale. Then we transformed the time series problem into supervised learning X and y matrix by shifting rows. Predicting next three months value is a multi step problem. We have historical observations (t-1,t-2,..t-n) forecast t, t+1 and t+2. Then we divided data into train and test set. We left the recently 20 months records as test set and other records as train set.

There are Vancouver East and Vancouver West with each has four types of benchmarks: residential all, detached, townhouse and apartment. Thus We prepared data and built 8 LSTM models for each of the benchmark. The predicted result can be fouond in our web UI property value trend page. The results and evaluations are further explained in the evaluation part.

In the second part, we merged tax reports from 2006 to 2018. Since the records before 2013 doesnt have the property type(whether its a land or strata property). When we merged tables, we based on the property PID in 2018 table. Then we calculated the average value of land property and strata property respectively from 2006 to 2018 per area. It is again a time series problem with 12 data points for each area. There are 26 areas have land property and 23 areas have strata property. Their price distribution is like in Fig. 3



(a) Land Property (b) Strata Property

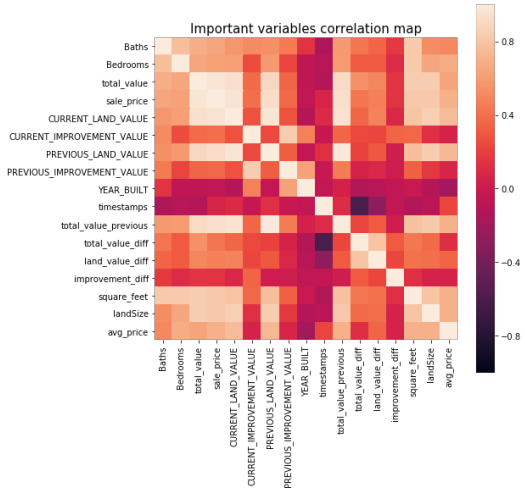
Fig. 3

For each area, we converted the time series problem into a supervised learning problem and applied gradient boosting regressor to it. There are 26 models for land property and 23 models for strata property. A grid search with cross validation is applied to find best model for each area. Due to the small size of data, the evaluation result wasn't stable. But most of the models can reach R^2 value above 0.8. The predicted result can be found in our web UI property value trends page.

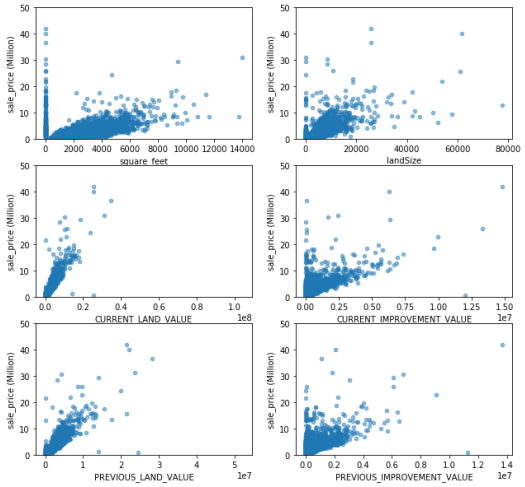
3.3.3 *Buyer Scenario*

To predict the purchase price, we chose to use the combination of the data from BC assessment and Vancouver tax report. We further visualized the relationships between these

numerical features and price using both correlation map and scatter plot Fig. 4. With the correlation map, we found that some features have a very strong linear correlation with the sale price (eg. square feet, land size, average sold price for different house types, and etc). To feed these features into the regression model, the two categorical features, subarea and the house type, were transformed into the vector with one hot encoding method. We tried 4 different regression models from scikit-learn package to predict the purchase price. They are linear regression, epsilon-support vector regression, k-nearest neighbors regression, and Gradient Boosting for regression. Moreover, we performed a 5-fold cross validation on these four models, and calculated the average of the R^2 scores. From our results, the linear regression model performs the best, which achieved an r square of 0.937. A further evalution of our model will be dicussed in Section 5.



(a) Spearman correlation map



(b) Some numerical features show linear correlation with the purchase price

Fig. 4: Investigation of correlation between each feature and the purchase price

3.4 Data Visualization and Web User Interface (UI)

We developed an web application to show predicted results to users. This application was built by using the python flask framework as the backend and the web server. Also, we used MongoDB as our database to store all the data we need in this application. The user interface was built with CSS/HTML, Javascript and Bootstrap. We used the Flask-GoogleMaps package to include Google Map. Some modifications were carried out to support the update of the neighbor information upon clicking the marker. To simplify the environment setup and avoid unexpected errors caused by different package versions, we configured to run our application inside a vagrant box. Therefore, users could just clone our repository and run command vagrant up inside the visualization folder to make the whole application work. Also, we deployed our application on AWS EC2 in order to make users easily access our application. Finally, we manually deployed a MongoDB server on AWS to support our application. The website URL is http://18.220.106.126:5000/.

4 Methodology

Here is a list of techinques we used in this project:

- Web scraping: Python selenium; Python requests
- Data cleaning, extraction, and integration: Python BeautifulSoup; Python pandas
- Data analysis, Data modeling: Python matplotlib; Python scikit-learn; Keras
- Visualization and Web UI: Python flask framework; MongoDB; Python pygeocoder (retrieving the latitude and longitude with address); Google map API (showing the request results on the Google map); leaflet; Chart.js, D3.js, AWS EC2

Why and how we apply these techniques are covered by Section 3.

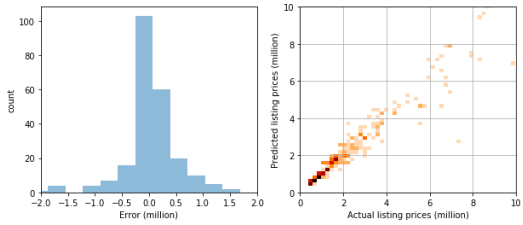
5 Evaluation

In terms of the functionality, we are very confident about our solution. Our product can serve almost all the roles in the housing market. Moreover, our solution tries to answer a very valuable question, whether the value of a property will increase or decrease during the following period of time. In terms of the accuracy of our models, we will illustrate why our predictions make sense in this section.

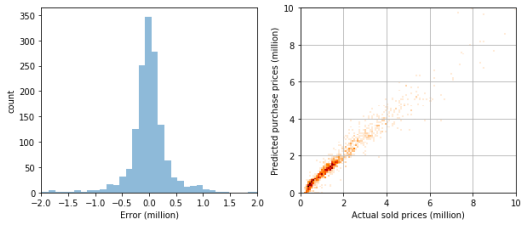
When we built the predictive models, we selected the model with highest R square. To further confirm our models make sense, we calculated the normalized root mean square error (NRMSE) by dividing the root mean square error (RMSE) with the difference between maximum and minimum values. Root Mean Square Error (RMSE) is the standard deviation of the prediction errors which measures how spread out the prediction errors are. A low RMSE means that the data points are concentrated around our model. Normalizing the RMSE facilitates the validation of models with different scales. Low values of NMRSE indicate small variance of prediction errors. For the listing price prediction, the model trained on the training data achieved R square of 0.964 and normalized RMSE of 0.037 on the test data (237 records). The error distribution is shown in Fig. 5a. The expected error is zero. The errors we produced satisfy the normal distribution. In addition, pcolormesh plot of predicted listing prices versus actual listing prices is

also shown. The slope is quite close to 1 which means our predictions are accurate.

For the purchase price prediction, our model achieved R square of 0.937 and normalized RMSE of 0.0103 on the test data (1447 records). With the same model evaluation approach, the error distribution graph and pcolormesh graph for purchase price prediction were also plotted (Fig. 5b)



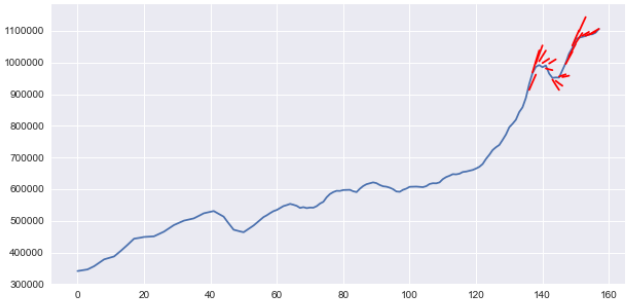
(a) Listing price prediction error distribution shows normal distribution. Predicted values are quite close to the true values.



(b) Purchase price prediction error distribution shows normal distribution. Predicted values are quite close to the true values.

Fig. 5: Our models perform well on test data

In terms of evaluating models in predicting future trends, a rolling-forecast scenario was used. Different from buyers and sellers model evaluation, RMSE is used to evaluate the model. Since we have 8 models, we tuned parameters to get each model to reach its lowest RMSE. The Predicted results of 'Vancouver East Residential All' is shown in Fig. 6 (plots of other models are in Appendix A). The blue line is the true value and red line is the predicted value. The more red line close to the blue line the more accurate the model is.



(a) Vancouver East Residential All

Fig. 6: The prediction almost fits the true value, especially in predicting t+1

6 Data Product

To make it easy for users to make use of our predictive models and to visualize the future trend of the Vancouver housing market, we built an interactive web application named VHMD. Based on our user scenarios, our web application has 4 pages, the main page, the seller page, the property value trends page, and the buyer page. The home page provides an overview of our tool and a brief description for the other three pages. The seller page allows users to get an estimated listing price for their properties as long as the detailed information of the property is given. Users can also compare their properties with the listing properties in the neighborhood to get the idea of how much their neighbors want to sell their properties for. The property value trends page allows users to get an overview pictures of each areas price trends. Users can learn if the property values in certain areas will increase or decrease. The buyer page allows users to get an estimated purchase price of the property with entered address. Users can also compare the property they want to buy with the sold properties in the neighborhood. At the bottom of the buyer page, the trend of the average sale price for the searched property type is displayed.

7 Lessons Learned

One important lesson we learned from this project is the experience of building the whole pipeline of the project. We need to define the problem we want to solve, collect useful data from different data sources, integrate the data, clean the data, extract the features from the data, build the regression models, evaluate the models, create an interactive web user interface to further demonstrate our findings. We put efforts to every single step of the whole pipeline. In addition, we realized the significance of data. Preparing structured data from scratch is painful. But decent data preparation will lead to high quality data model.

8 Summary

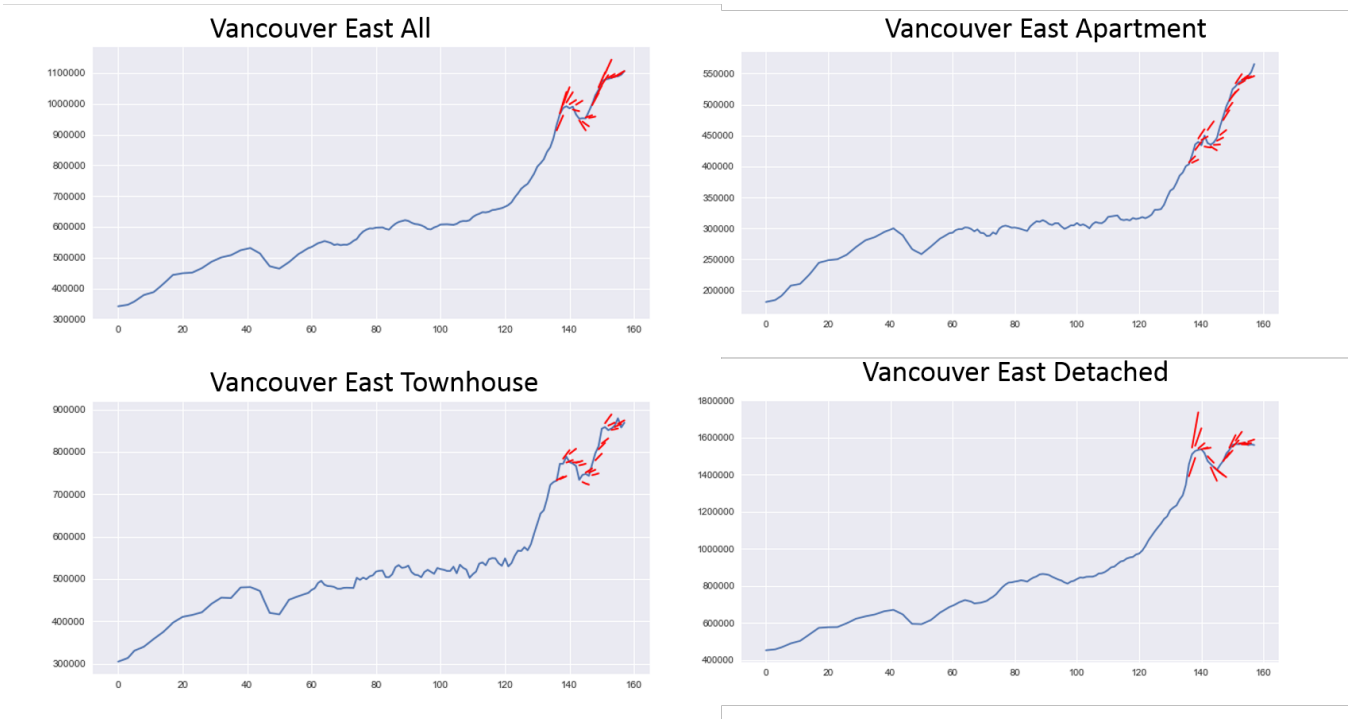
In this project, we built high quality machine learning models based on the crawled data to predict property current value and future market trends. The web based tool (VHMD) we created achieved three goals, predicting estimated listing price for sellers, predicting estimated purchasing price for buyers and showing the future market trends in different areas for different types of properties. Generally speaking, the values of the properties in Vancouver will remain stable for the following three months.

References

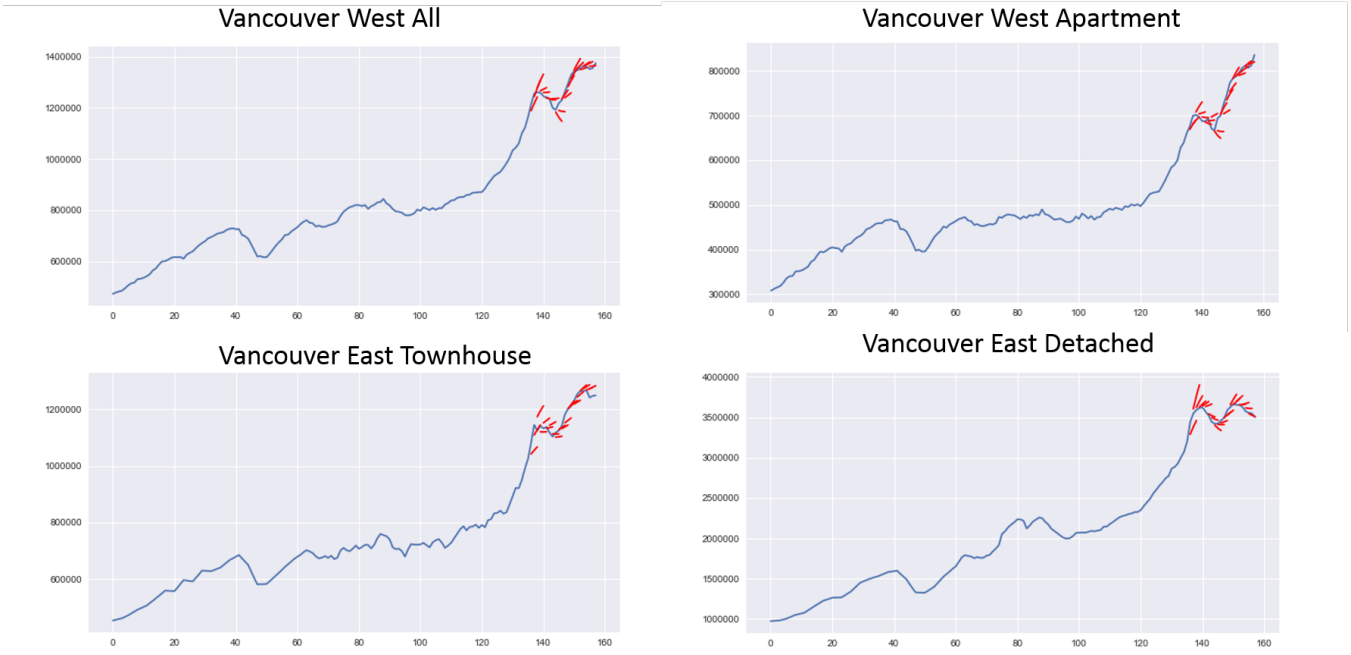
[1] Jerome H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". The Annals of Statistics (2001) Vol. 29, No. 5, pp. 1189-1232

[2] Sepp Hochreiter; Jrgen Schmidhuber. "Long short-term memory". Neural Computation (1997) Vol. 9, No. 8, pp. 17351780

APPENDIX A



(a) Vancouver East Predicted vs True trends



(b) Vancouver West Predicted vs True trends