Metro Vancouver Housing Market Analysis and Prediction

GitLink: https://csil-git1.cs.surrey.sfu.ca/nmisra/eigentum/

Team: EIGENTUM

Krishna Chaitanya Gopaluni

Nitin Misra

Harish Bhargav Dasika

Manjur Rahaman

Content

Motivation and Background

Predicting the house price serves the need for two group of people, buyers and sellers. Most people interested in the housing market are intrinsically motivated by the timing and both short and long term value of a property. Since historically Vancouver has always been in a housing bubble. Potential buyers take the current increasing price trends of housing for granted to invest. But the prices may suddenly fall and it takes really long time to get Return On Investment. In this project we aim to build a system that would aid potential user about the existence and nature of housing bubble in an area, an estimation of housing price based on a number of features and a future property value prediction for a number of different property type in east and west Vancouver.

Problem Statement

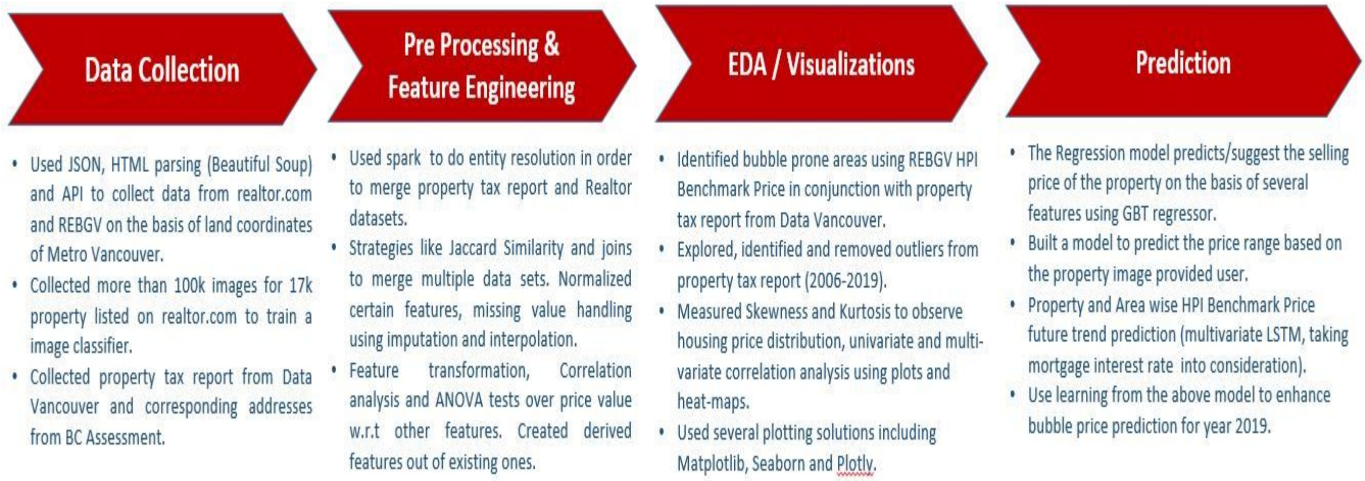
Identifying housing bubble is hard. It is even harder for a place like Vancouver and BC since both these places have historically been showing strong resistant to housing bubble[1]. As far as we know currently no real estate solution has attempted this problem successfully. Another integral part of our project was to give an accurate estimation of both the current and future price trend of an property. Predicting the future trend is a very difficult problem since time series models are rarely good enough in real life scenario.The nature and availability of different data sources added to the difficulty of the problem since it was not obvious how to collect, merge all these different data from a number of different sources.

We spent a significant amount of time on collecting, cleaning and merging data. And last but not least we trained an image classifier to classify house price range based on image data. The goal was to make a comparison of estimation between our regression model prediction and image classifier prediction and serve different user depending on the context of the use case. Our project goal can be described in four distinct ways.

1. Identifying Bubble Prone Areas in Metro Vancouver.
2. Predicting the housing price based on current trends.
3. Predicting HPI Benchmark Prices future trends.
4. Predicting property price range based on property image.

In the end, we hope our solution would help a potential buyer in warning about bubble-prone areas and he will be able to make informed decisions based on future trend prediction. A seller will know the current value of his house based on certain features.

Data Science Pipeline



Data Collection

- Used JSON, HTML parsing (Beautiful Soup) and API to collect data from realtor.com and REBGV on the basis of land coordinates of Metro Vancouver.
- Collected more than 100k images for 17k property listed on realtor.com to train a image classifier.
- Collected property tax report from Data Vancouver and corresponding addresses from BC Assessment.

Pre Processing & Feature Engineering

- Used spark to do entity resolution in order to merge property tax report and Realtor datasets.
- Strategies like Jaccard Similarity and joins to merge multiple data sets. Normalized certain features, missing value handling using imputation and interpolation.
- Feature transformation, Correlation analysis and ANOVA tests over price value w.r.t other features. Created derived features out of existing ones.

EDA / Visualizations

- Identified bubble prone areas using REBGV HPI Benchmark Price in conjunction with property tax report from Data Vancouver.
- Explored, identified and removed outliers from property tax report (2006-2019).
- Measured Skewness and Kurtosis to observe housing price distribution, univariate and multi-variate correlation analysis using plots and heat-maps.
- Used several plotting solutions including Matplotlib, Seaborn and Plotly.

Prediction

- The Regression model predicts/suggest the selling price of the property on the basis of several features using GBT regressor.
- Built a model to predict the price range based on the property image provided user.
- Property and Area wise HPI Benchmark Price future trend prediction (multivariate LSTM, taking mortgage interest rate into consideration).
- Use learning from the above model to enhance bubble price prediction for year 2019.

The above figure portrays our data science pipeline. A step by step description of each of the components in our pipeline is described below.

Data Collection

We started exploring around out problem statement and the solution we wanted to give and found out that we can't use any prepared data to solve the business problem. We started to search for the property listing websites first and shortlisted Realtor.ca and Rew.ca for getting the data properties listed for selling. Based on the popularity and the huge database of all the active property listing in Metro Vancouver we selected Realtor.ca as our primary data source for all the properties listed in Metro Vancouver. We also decided to collect the data from Data Vancouver which keeps all the historical property tax report for the Vancouver and Greater Vancouver area.

We needed a benchmark index to solve the problem of bubble analysis thus we chose HPI Property Benchmark Price (HPI-PBP). HPI-PBP is calculated using multivariate regression analysis, a commonly used statistical technique for all the properties listed in an area. HPI-PBP in British Columbia is managed by two agencies namely CREA - Canada Real Estate Agency (crea.ca) and the REBGV - Real Estate Board of Greater Vancouver (rebgv.org). Both of which uses different set of criteria to calculate the HPI-PBP in Metro Vancouver and we used both of the data source from year 2006-2019.

Using the property tax report from the Data Vancouver we collected the street name of the properties from the BC Assessment with the help of a open REST API, later we used the street name to merge with the realtor properties 2019 data set. We also collected the mortgage interest rate from Bank of Canada for past 15 years (2006-2019).

Realtor Property Listing

We collected the properties listed for sale on realtor.ca through use of some open REST API APIs. It was hard to get the data set for Metro Vancouver, since the website uses land coordinates to fetch the active property listings for sale. Using the set of different land coordinates we tried to cover as much area as possible, so we don't lose any listings in the process of fetching the data, later we dropped the duplicated data on the basis of the MLS number.

Before removal of duplicate we fetch approximately 86K records, later after removal of duplicates reduced this number to approximately 17K. We didn't stop there and discovered two new set of data for each property which could enhance our predictions and results, so we started parsing data from the details page of each listing using HTML parsing (Beautiful Soup) and also found a open REST API to fetch information about nearby surrounding (proximity score out of 5 for daily necessities like schools, grocery shop etc.) of the properties. We also fetched property listing images so we can later use it to solve parts of our problem statement. Whole process of discovering resources for realtor and fetching the data took more than 4 weeks. At the end we had 17K active property listing on realtor with more than 100K images for property listing.

Vancouver Property Tax Report (2006-2019)

Our next data source was Data Vancouver which keeps all the data set related to property tax in whole Vancouver area. The property tax reports are prepared and managed by BC Assessment and contains the information about the property current land value, improvement value, property tax information for the past year, street name along with many more important features for each property.

Property Benchmark Price (2006-2019)

As discussed in the introduction of this section we used two Property Benchmark Price from REBGV and CREA.

- REBGV: We utilised the HPI-PBP matrix available with REBGV to get the HPI-PBP for different types of properties for all the areas in British Columbia using PDF and HTML parsing.
- CREA: We utilised an existing document containing HPI-PBP information for all the areas in British Columbia without any specialised technique.

Mortgage rates from BOC (2006-2019)

Bank of Canada provides a document which contains the daily mortgage percentages for different types of plans (1y-fixed, 1y-bond….7y-bond etc) which identifies the time period in which mortgage needs to be paid back. Among all the choices we had we chose 5y-fixed-posted for our calculations since it's the most common plan people chooses to pay-off their property mortgages.

Data Cleaning, Data Extraction and Data Integration

Vancouver Property Tax Report (2006-2019)

Vancouver property tax property report consist of the all the properties in Vancouver with previous year tax information and many other important fields of the property. We merged all the tax reports and merged it with the 2019 data set and got around 209K records. We also required address details of this 209K records (since it's not available due to privacy issues), to fetch the address we exploited BC Assessment's open REST API which provides the address details on the basis of PID available in the tax report. We used several techniques to handle null values like filling them with mean and interpolation, we also dealt with the inconsistent data format which were important but not in required format appropriate for machine learning models processing, so we normalised them using different technique. We also calculated some derived fields like the House_Price which represents the house price on the basis of CURRENT_LAND_VALUE and CURRENT_IMPROVEMENT_VALUE. After integration and cleaning we selected 20 features on the basis of correlation analysis (box plots and heatmaps) by mostly following statistical feature selection techniques.

Realtor Properties Data

The data we got from luckily in JSON format because we invested time and identified how realtor is fetching the data on their website and we replicated the same request to get the list of properties. The data from the details page of each property was in HTML format so we decided to use Beautiful Soup python library to get the most relevant details from the details page. For getting the daily necessities' proximity scores we discovered and used an open REST API from realtor which returned us data in JSON format. We used data from the details page and the proximity score API and merged them into the original list of properties from realtor. We fetched all the property images which took most of the time because we had to download more than 100K images for more than 17K properties.

After data collection from realtor we found 40 features in which some were not appropriately formatted for the analysis and some of them were not even required, so we had to discard several features and apply normalisation techniques on the data. Later we need to merge the realtor properties data with the data we collected in previous step, so as do further analysis and build machine learning models.

Since there were no common column to join the two datasets, we decided use Entity Resolution technique (Jaccard Similarity) on address field. We used property address from the realtor data set and the address from Vancouver property tax report and bais on Jaccard Score we filtered the realtor dataset and merged the Vancouver tax report and ended up having 82 columns among which we dropped several columns and selected 28 numerical and 12 categorical features.

HPI Property Benchmark Price Data (2006-2019)

We used HPI-PBP widely for the bubble calculation and analysis. We used two features from CREA i.e the date and the Composite Benchmark Price on the date and then grouped data on year to have have a media benchmark price for each year from 2006-2019. Using REBGV data we used Area, Benchmark Price and Property Price then grouped all the data on year and got median Composite Benchmark Price for all 4 type of properties in Vancouver East/West area, which greatly improved the property bubble analysis.

Bubble Analysis

Vancouver as per past statistics has always been under bubble based on the statistics. This has motivated us to come up with a strategy to identify the Bubble prone Areas. Which was achieved with the help of BC Assessment Dataset and HPI index data (by REBGV), which explained below.

BC Assessment (Dataset),which develops and maintains the real property assessment throughout the British Columbia. Features such as 'CURRENT_LAND_VALUE' and 'CURRENT_IMPROVEMENT_VALUE' have been considered and HOUSING PRICE has been calculated ('CURRENT_LAND_VALUE' + 'CURRENT_IMPROVEMENT_VALUE' = 'HOUSE_PRICE'). And, using the dataset we have got the Median of the housing prices with respect to each area over the years (15 years) have been taken into account.

On the other side, we have taken the HPI index Dataset. There are two Organization which calculate HPI index monthly, one was by REBGV and the Other by was CREA. After good amount of study and understand, based on the diversity of the factors considered in REBGV calculation, we found that it was far more fetching than CREA's (more generalized) methodology.

Having said that, the HPI Index obtained from REBGV was given more weightage. From this we are considering the "Composite" benchmark price which considers all types of the houses while calculating the benchmark price. As said, HPI dataset is consisting of Composite Benchmark price with respect to each area in Metro Vancouver for the past 15 years from (2006 to 2018).

After having a calculated analysis on the Housing Price, we have noticed potential areas whose house price abnormally high. Hence, we with the concept of IQR (Interquartile range), have eliminated Outliers which finally lead us to a proper set of data. Later on, faced issue with the scale of records. Hence, with the help of Scikit-learn Scaler, we had to scale the Housing price with a ranging between the Min.

and Max. values of the composite benchmark this has helped in getting the data into proper shape.

Now moving to the actual Bubble Detection part, with the help of above shaped Housing price set and Composite Benchmark price set, we have calculated an area wise percentage change, for all records(yearly) i.e. Percent Change (PC) = ((HOUSE PRICE – Composite Benchmark) / Composite Benchmark) * 100. Upon doing the above, we have considered two Conditions (Filtering Criteria) to decide our Bubble prone area.

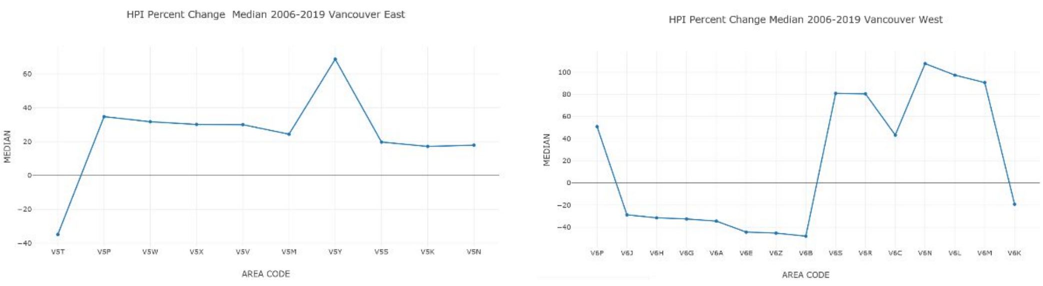
Filter1(*Threshold Percentage)*: The Obtained Percent Change (PC) has been taken as the base for our bubble estimation, all the records which are falling outside the percentage Threshold (-10 < PC < 10) are considered as '*Vulnerable*' to bubble.

Filter2(*Frequency*): From the above results, all the areas which are showing up 7 or more times in the span of 15 years are marked as '*Highly Vulnerable*' areas.

FIG[1] Shows the area's HPI Percent Change Trends over years, (Highly Vulnerable Areas),Where [1A] is for East and [1B] for West .

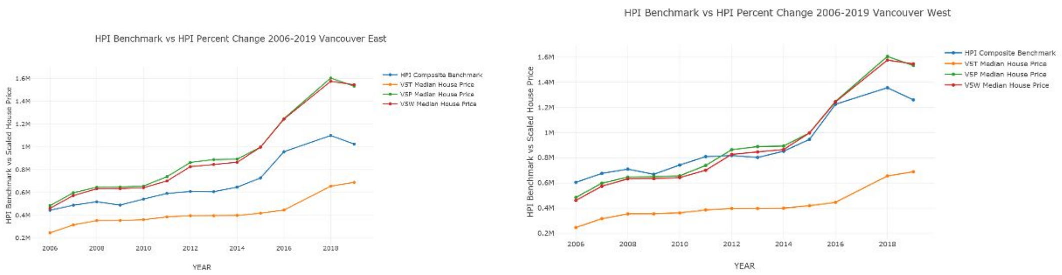
FIG[2] Shows comparison between the HPI Benchmark and Median House price (2006-2019), where[2A] is for East Vancouver and [2B] for West Vancouver.

Finally, A sort was performed considering the percent change and frequency of areas (showing up as Highly Vulnerable), and the top records have been considered. All the Top 5 areas have been concluded as the "***BUBBLE PRONE AREAS***."
Above mentioned calculation was performed among all the #Areas in Metro Vancouver (Both West and East).
FIG[3] Shows House price fluctuations over the years for the BUBBLE PRONE Areas(Top 5 areas), where [3A] is for East Vancouver and [3B] is for West Vancouver.



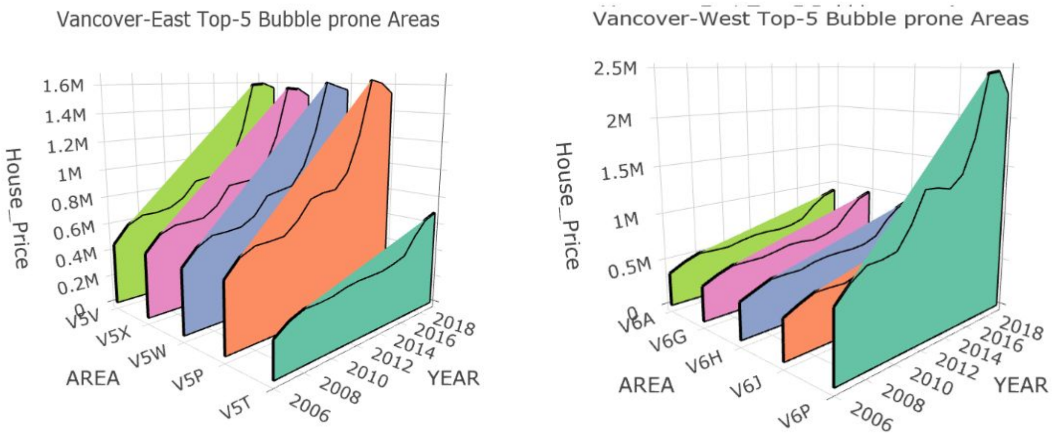
[1A]-East [1B]-West

FIG[1] HPI Percent Change Trends over the years(Vulnerable Areas)



[2A]-East [2B]-West

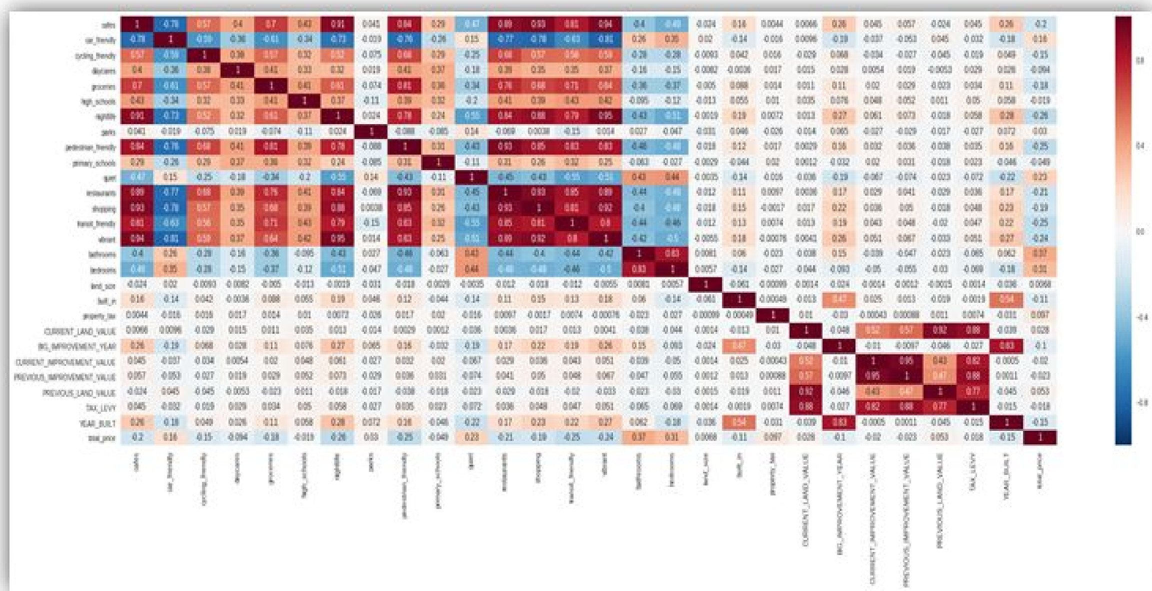
FIG[2] HPI Benchmark and Median House price (2006-2019)



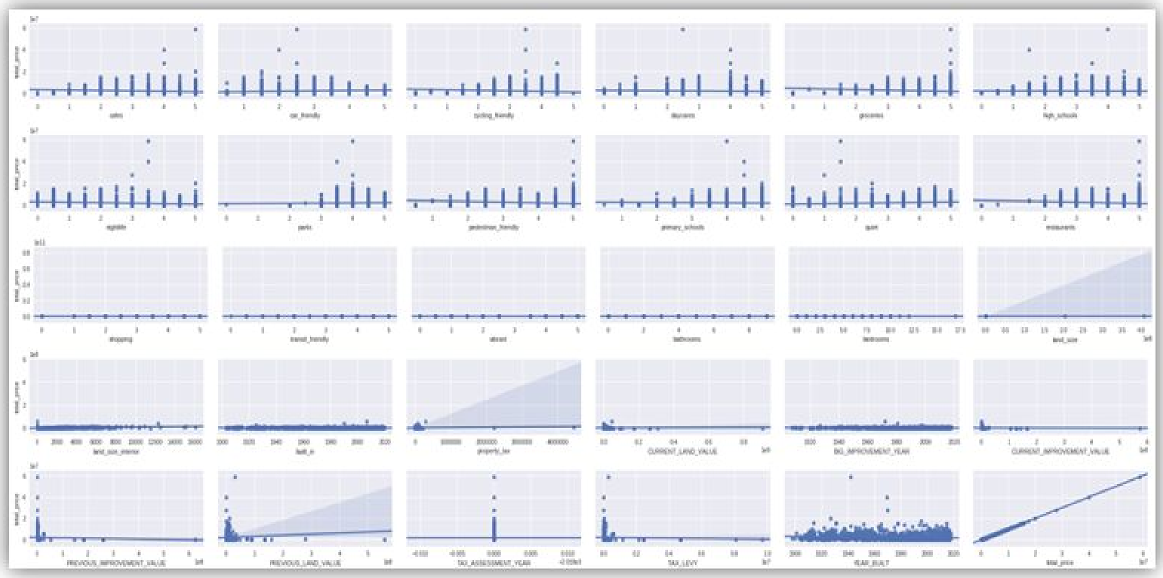
[3A]-East [3B]-West

FIG[3] House Price fluctuations for top 5 Bubble Prone Areas over the years

Regression Model For Predicting House Selling price



FIG[4] Heat map representing how numerical features are correlated to house price



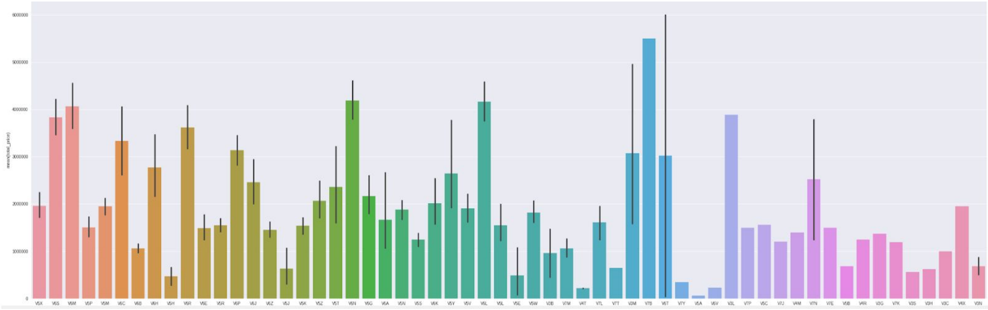
FIG[5] Correlation plot for different features vs total price

Datasets we used for regression are the scraped data from Realtor containing the details of current properties on sale and 2019 property tax report from Data Vancouver.

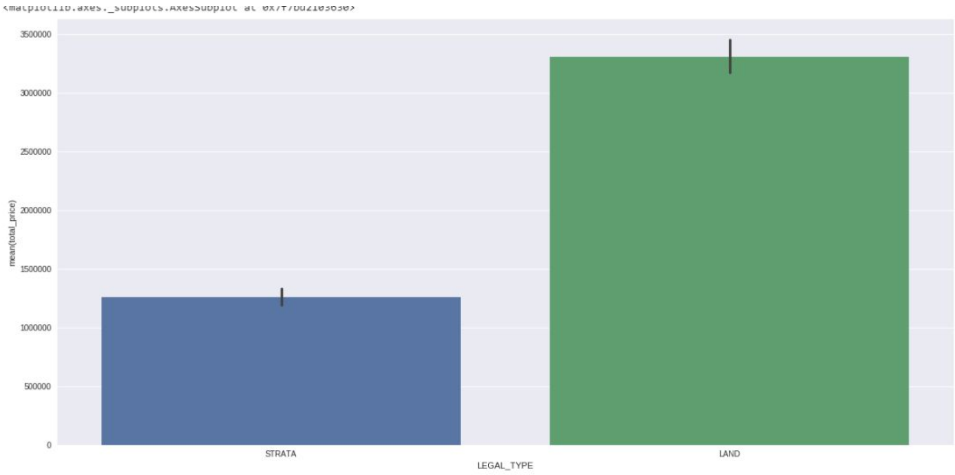
Using PID from 2019 property tax report we scrapped address from BC assessment and implemented simplified Entity Resolution using Jaccard Similarity based on street name to combine Realtor data with property tax data.

Pre processing involves removal of string from numerical features, filling the null values using imputation and interpolation, Normalizing the data. We have used the domain knowledge to create new features like total price of the property using the previous and current land improvement values.

We built a regression model using 28 features (numerical, categorical) using GBT Regressor to predict selling price on this data and achieved a good r^2 of approx. -0.70. Grid search with 5 fold cross validation was applied to find the best parameters for GBT and random forest.Plots on the top shows *pearson* correlation analysis of numerical variables while the below ones are categorical variables (Area Code, Legal Property Type & Property Type) plotted over price value.



FIG[6] Price ranges plotted over the Area Codes



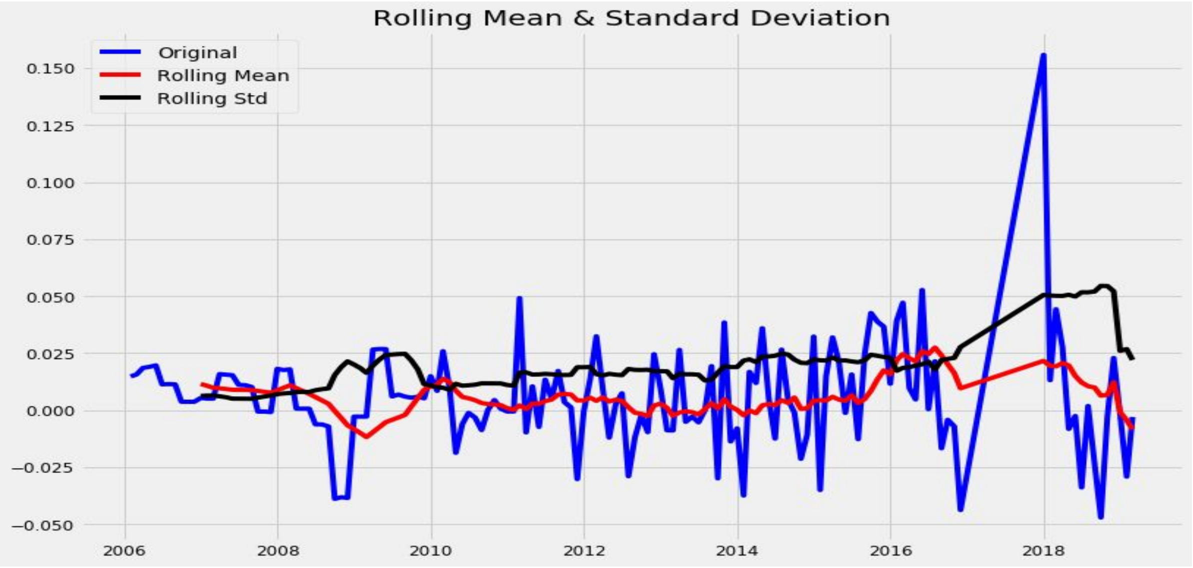
FIG[7] Property type vs price range (“land”(green)- owned by government, STRATA(blue)- privately owned)

Using images from the Realtor.ca to predict the price ranges

We have scrapped realtor images and saved each property images under MLS number. Later we have tried standardizing the data set by placing the images into 100 classes, created based on the price ranges.Created training, validation, testing folders. We have trained a sequential model(2D convolution layer, 5 hidden layer, softmax and relu activation) on keras to detect the price range based on the property image. Our goal was to make a better estimation based on both the prediction we get from our regression model and the house price range prediction we get from our image classifier. Although we do have a model working, we found this problem to be very difficult to solve and needs a lot of time and attention. In Future we hope to improve our classifier significantly for better prediction.

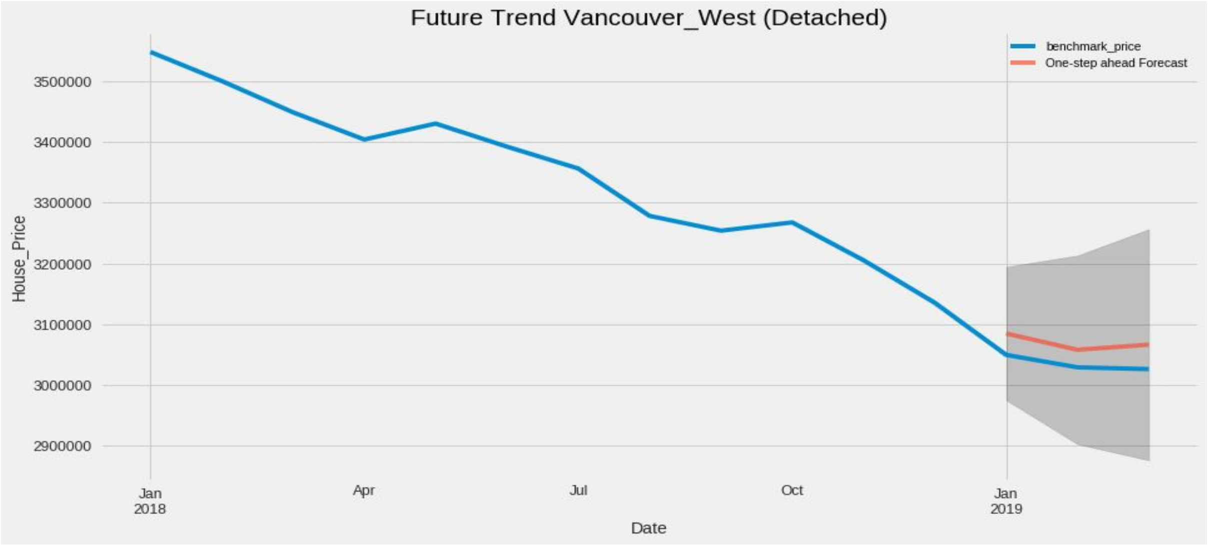
Predicting Future Trend

We predict future benchmark price using REBGB dataset. The dataset contains benchmark price from 2006-01-01 to 2019-01-03 area and property wise. For each of Vancouver east and west zone’s property type(apartment, town hall etc) we did time series analysis to predict the future trend. We converted the time series problem into supervised learning for by shifting row. This allows us to use t-1 observation(at timestamps t-1) to predict the value at t(matrix, x and y). We measure the stationary properties of our dataset using statsmodel adfuller which uses dicky fuller test. We do a log transformation and then take a shifted difference to make sure that the dataset is free of trend and seasonality and meets our threshold in dicky fuller test.

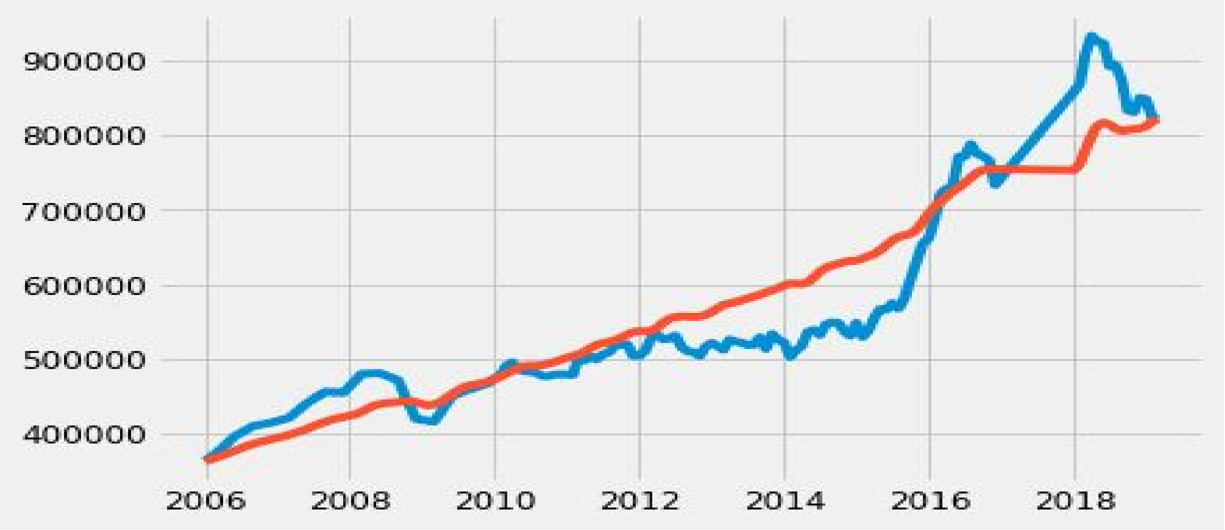


FIG[8] displays the rolling mean and rolling standard deviation after the log transformation and shifting difference.

As our time series model we used multivariate LSTM and SARIMAX. Since LSTM requires scaling(1,-1), we used minmaxscaler of scikit learn to transform our data set and later transform it back during the future prediction. We split our dataset into training(.7) and testing(.3). Since, we care about how far our prediction deviates from the optimal result we used RMSE as our evaluation metrics.



FIG[9] Displays the future trend predicted by the model for Vancouver west detached property type.



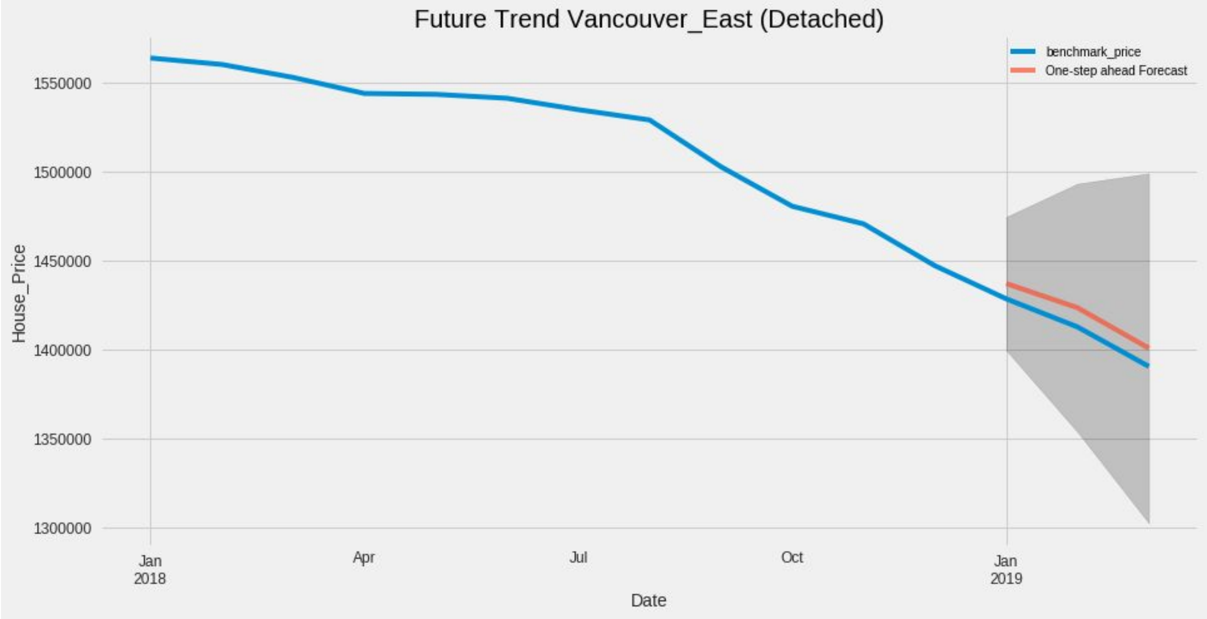
FIG[10] Future Trend Prediction for, Vancouver east-Townhouse

As we had few data points to train our model on for a number of property type, we used SARIMAX for those properties with fewer data points. We apply grid search to find the optimal parameters for our data combinations. We forecast for the next 10 months and save that result into CSV. Figure-10 shows the future trend prediction for Vancouver east townhouse.

Evaluation

Predicting housing bubble on the basis of HPI makes sense in our context as HPI itself is built on a rigorous multivariate regression model[2]. HPI measures housing price changes over time by taking factors into account. Each of this factors contribute to the house value, HPI not only takes fundamental factors such as location, year_built etc takes into account but also it considers Socio-demographic attributes, consumer income level, neighbourhood characteristics (proximity factors) etc in the final calculation of a house price. With the consideration that HPI does give us an accurate estimation of house price change over time, we are confident that our housing bubble calculation captures the bubbles prone areas correctly. A description of how the housing bubble is calculated is given on the section titled "**BUBBLE Calculation/Analysis".**

Before doing housing prediction using our regression model we used a log transformation to make sure that our distribution resembles that of a normal distribution(also fixing the skewness)**.** Since RMSE in our regression model indicates how much our prediction deviates from the actual, our RMSE. The low RMSE and r**2 we have achieved gives us reasonable estimation. In Order to further justify our model we have calculated NRMSE, RMSE/y_max()-y_min(). Normalizing the RMSE facilitates the comparison between datasets or models with different scales. NRMSE is often expressed as a percentage, where lower values indicate less residual variance. Our NRMSE value is estimated to be (.039). Which confirms the reasonable estimation as low variance in NRMSE indicates small variance in prediction error.



FIG[11] Future Trend Prediction for, Vancouver east-detached

For our future benchmark price prediction, we are again using RMSE as our evaluation metrics. We spent a significant time on parameters turning for the optimal RMSE we could get. We used rolling forecast to evaluate the performance of our model. Figure-11 shows the trend prediction captured by our model for Vancouver east detached property type.

Not all of our models for all each property type was able to capture and predict the future trend well. We attribute that fact to the unavailability of the data.

Data Product

Our data product is a collection of python notebook designed to provide a potential user with the solution of a number of problems related to real estate market. We focused on providing a solution for detecting housing bubble since we considered that problem to be noble. We attacked the problem of house price predicting from a different angle by utilizing image data.

Due to time constraint we could not build an actual product out of it but we believe we have everything in place to create a compelling web application.

Lessons Learnt

Predicting the future trend based on the present value is very hard. If the data does not have underlying hidden structure its very hard for time series model to learn non linearity in the data. In this project we have tried to build a data science project from scratch. Figuring out what, where and how to get the data for our project took 80% of our time which seems to support the notion that in real world data scientists spend majority of their time collecting and cleaning data. Building a classifier to extract features from the images to classify house price range is a daunting task that need more time and attention. We hope to get back to this problem in future. For bubble analysis we did not have a reference to look into, we hope our contribution for that particular problem would be helpful to others. in future we hope to combine the future benchmark prediction with our bubble analysis to calculate future bubble prone areas.

Summary

In this project we have collected, cleaned data from multiple sources to analyze the housing bubble for metro Vancouver and built machine learning models to provide an accurate estimate of current and future housing price trend for different housing property type in metro Vancouver. Our model shows a difference in pricing trend for various property type in Vancouver east and west zone. We also show the most bubble prone areas in metro Vancouver.

References

[1]Nobel laureate sees B.C. and Vancouver as ‘housing bubble resistant(https://www.whichmortgage.ca/article/nobel-laureate-sees-b-c--and-Vancouver-as-housing-bubble-resistant-185567.aspx)

[2]https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/

[3]https://www.vancouverisawesome.com/2018/10/01/vancouver-real-estate-bubble-inflated-world/