SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Metro Vancouver Housing Market Analysis & Prediction

Team : *Krishna, Nitin, Harish, Manjur*

MOTIVATION & GOALS

Vancouver is always in the bubble. Potential buyers take the current increasing price trends for granted to invest. But the prices may suddenly fall and it takes really long time to get Return On Investment. Keeping this in mind we have come up with the following goals.

- **Identifying Bubble Prone Areas in Metro Vancouver.**
- **Predicting the housing price based on current trends.**
- **Predicting HPI Benchmark Prices future trends.**

✓ In the end, the project can 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.

DATASETS

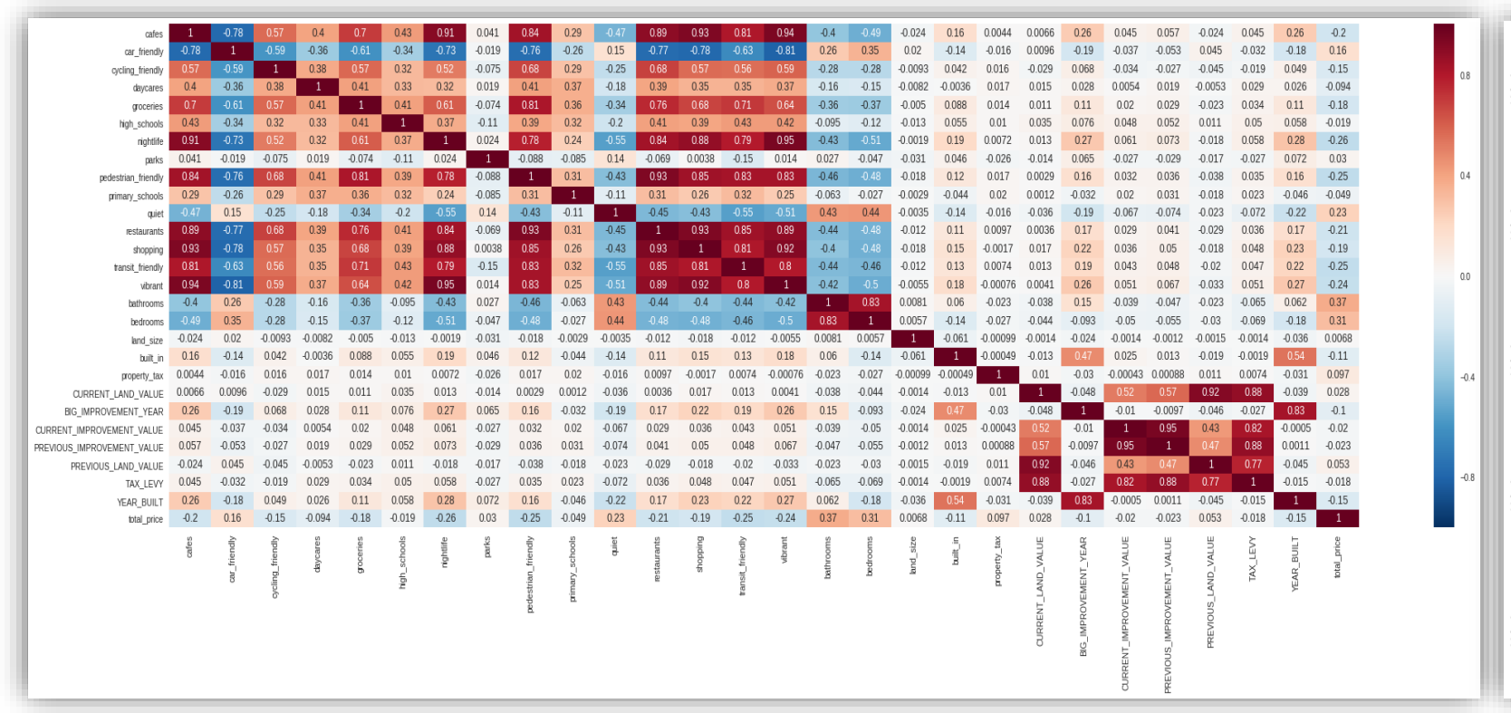
REALTOR: Biggest real estate listing agency in Canada
No. Records – More than 16K records
Features Used – 28 odd features
IMAGE data of 17k properties - 100K

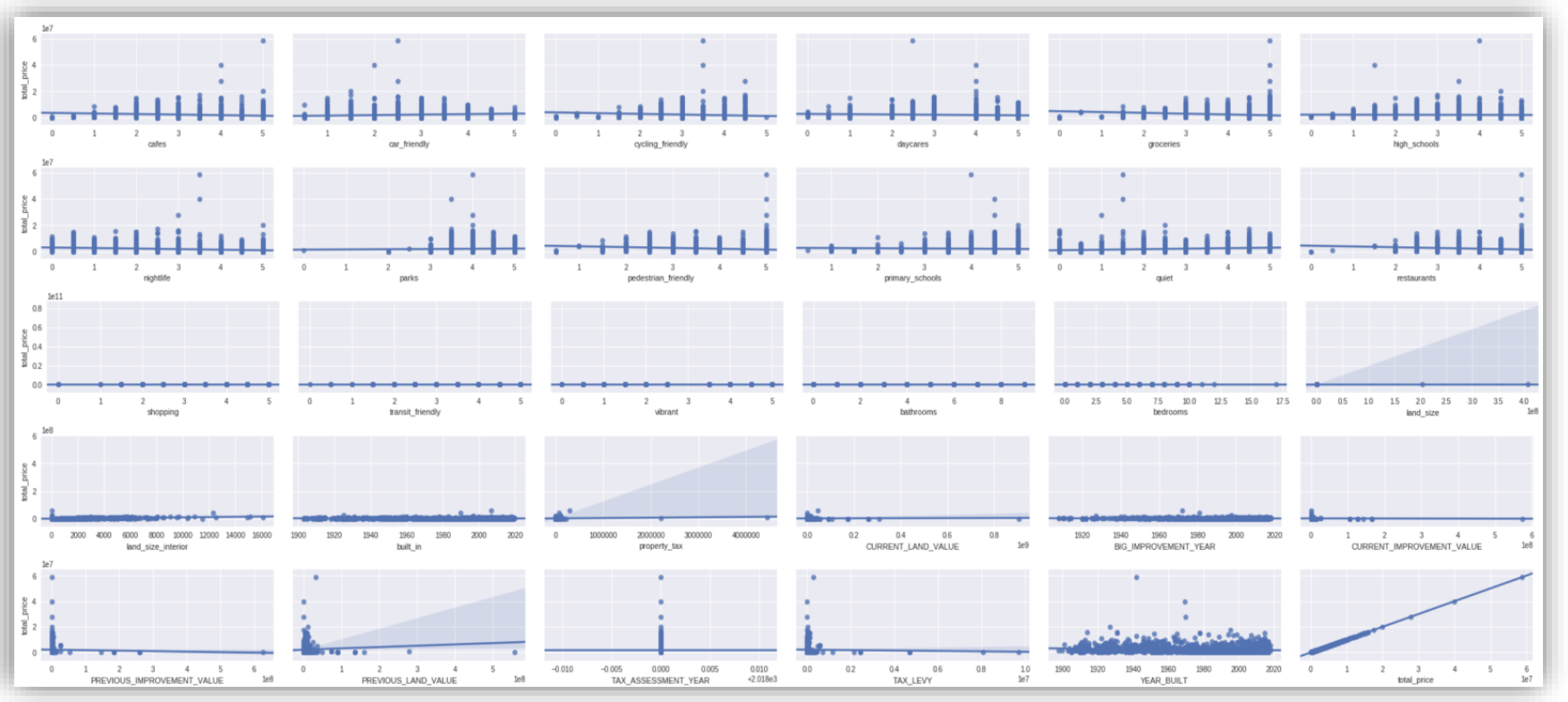
REBGV - HPI Index: House Price Index provided by Real State Board of Greater Vancouver
No. Records – 1175 records
Important Feature – Benchmark Price

BOC (Bank Of Canada): Property mortgage rates Information has been taken into consideration from the past 15years

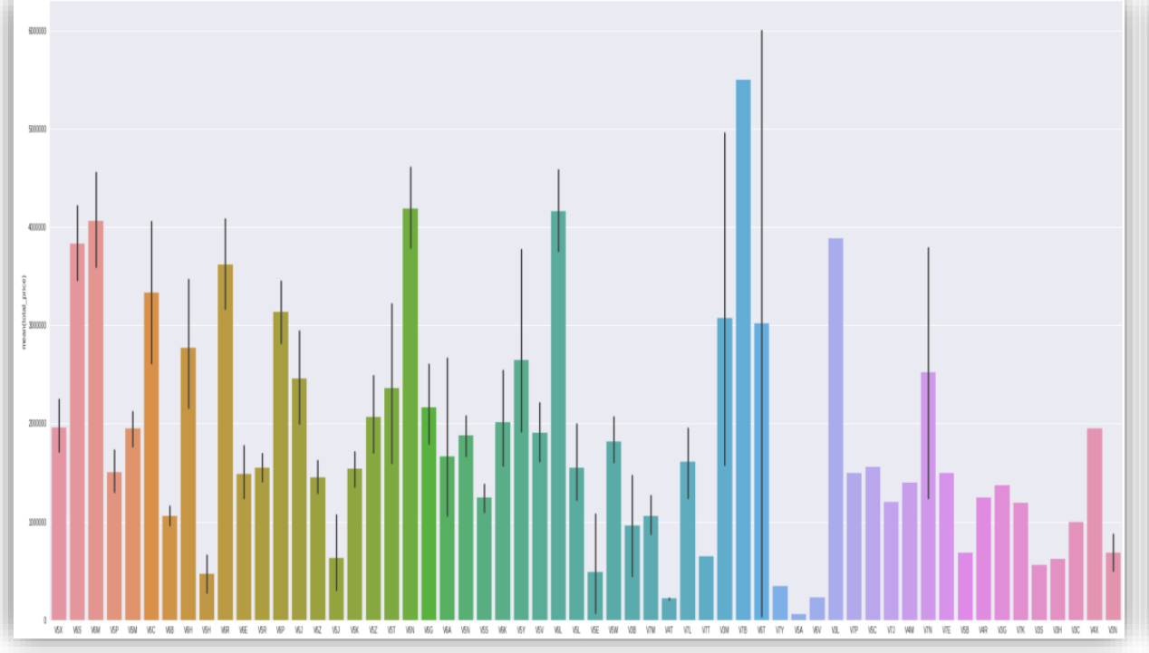
BC Assessment: develops and maintains real property assessments throughout British Columbia,
No. Records – More than 200K records
Important Feature – Addresses

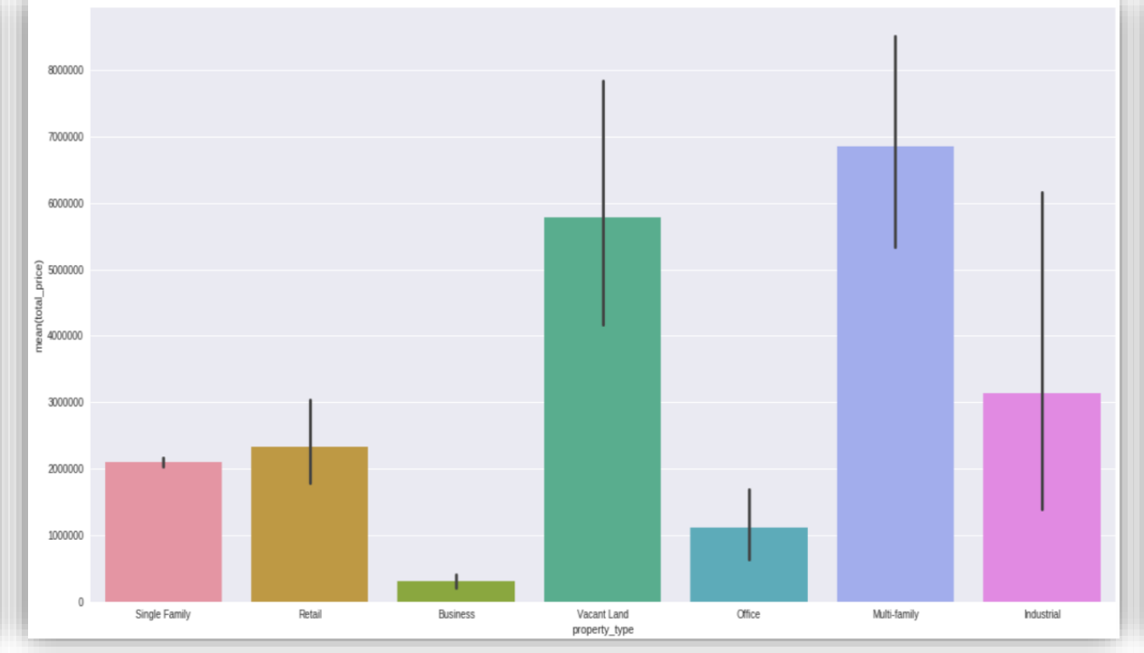
PREDICTING THE HOUSE SELLING PRICE

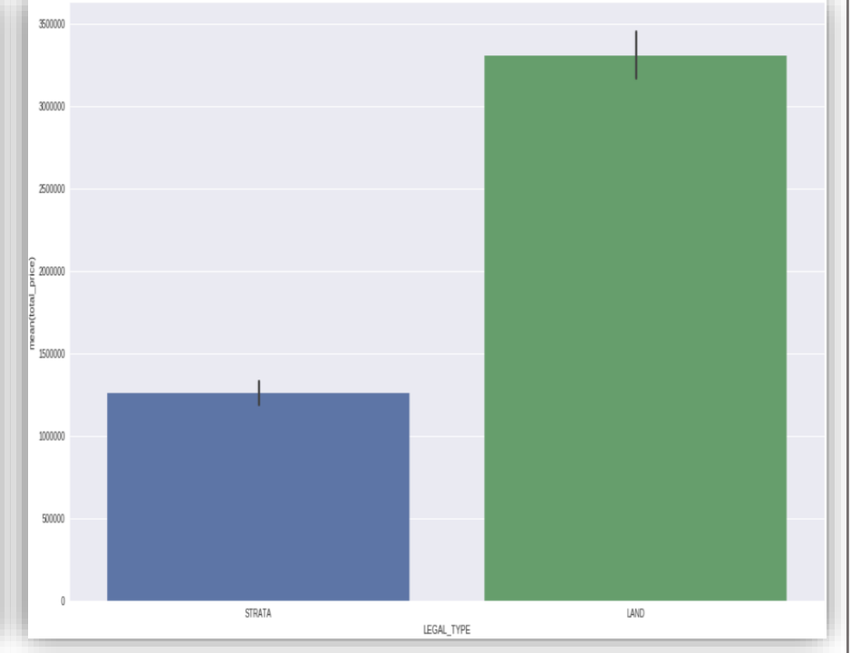




- We scraped the data of properties on sale from Realtor, and 2019 property tax report from Data Vancouver. Using PID from 2019 property tax report we scrapped address from BC assessment and implemented simplified Entity Resolution using Jaccard Similarity based on street name to combine Realtor data with property tax data.
- 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.
- Plots on the top shows correlation analysis of numerical variables while the below ones are categorical variables (Area Code, Legal Property Type & Property Type) plotted over price value.







PROJECT 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.

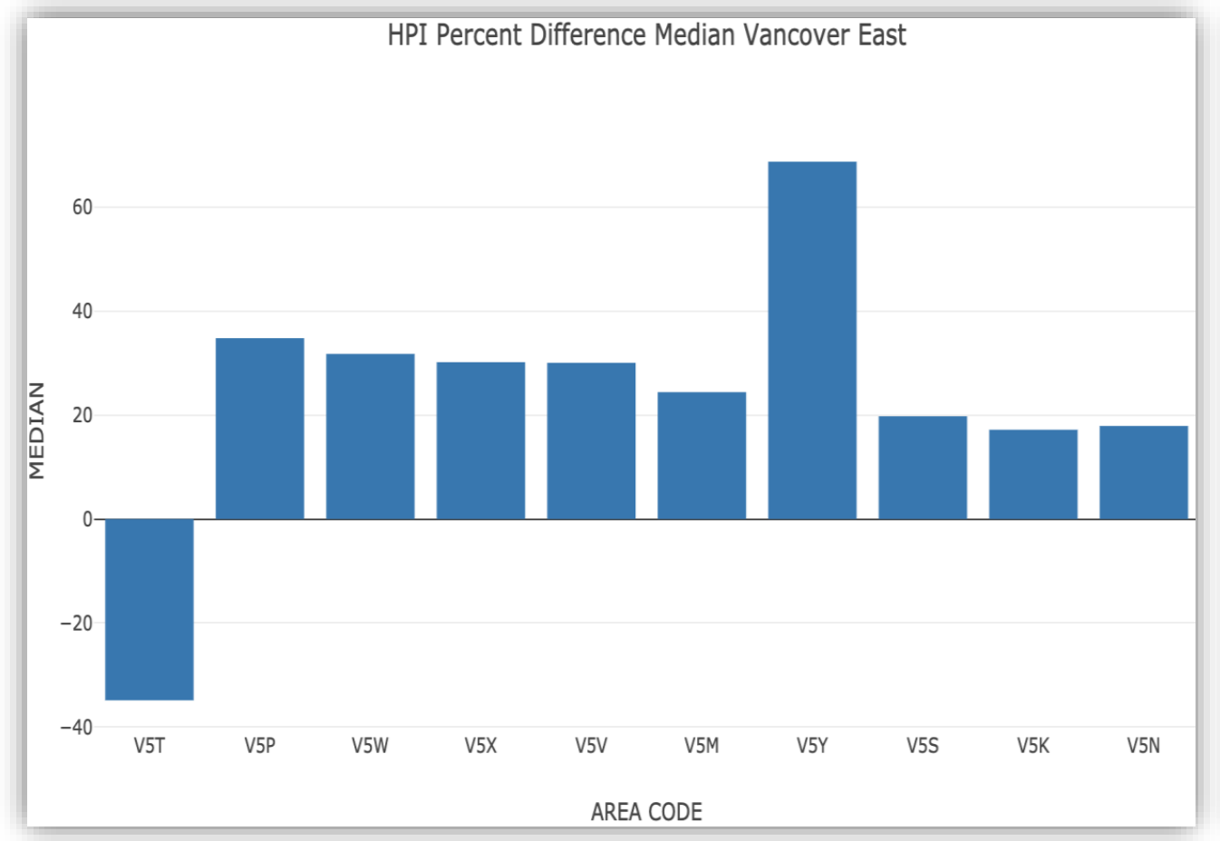
EDA / Visualizations

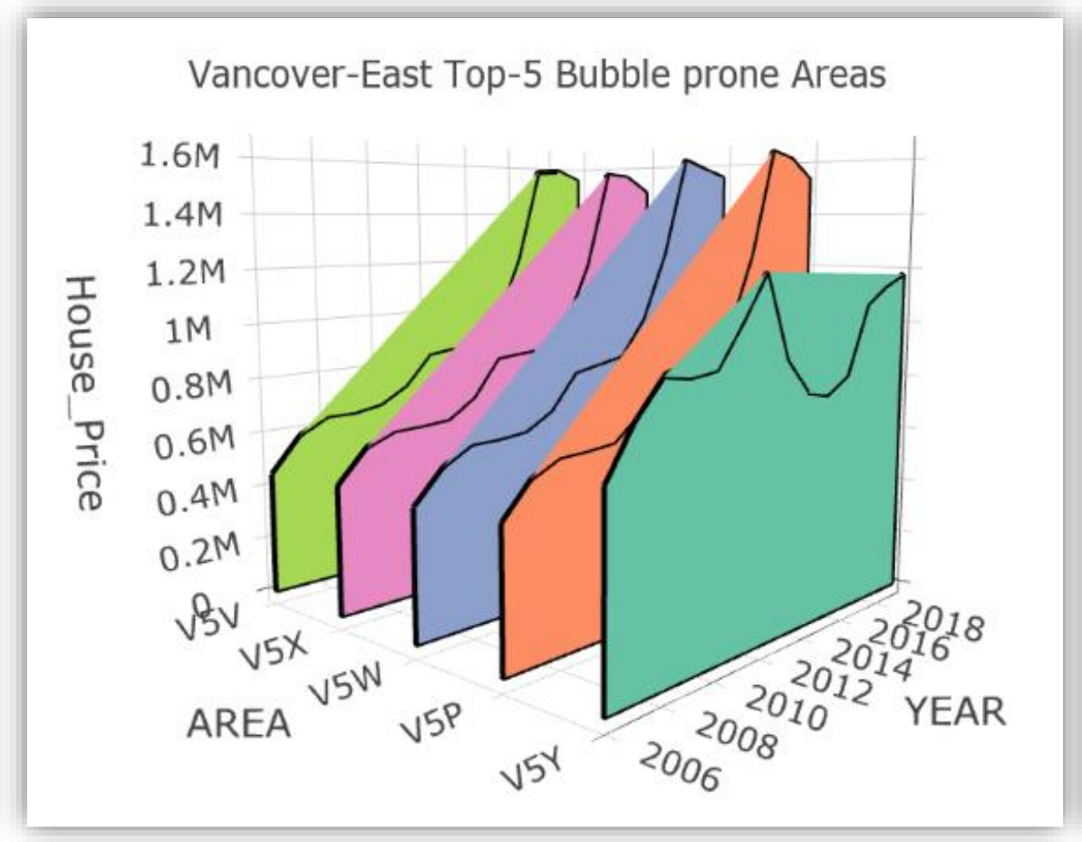
- Identified bubble prone areas using REBGV HPI Benchmark Price in conjunction with property tax report from Data Vancouver.
- Explored, identified and removed outliers from property tax report (2006-2019).
- Measured Skewness and Kurtosis to observe housing price distribution, univariate and multi-variate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

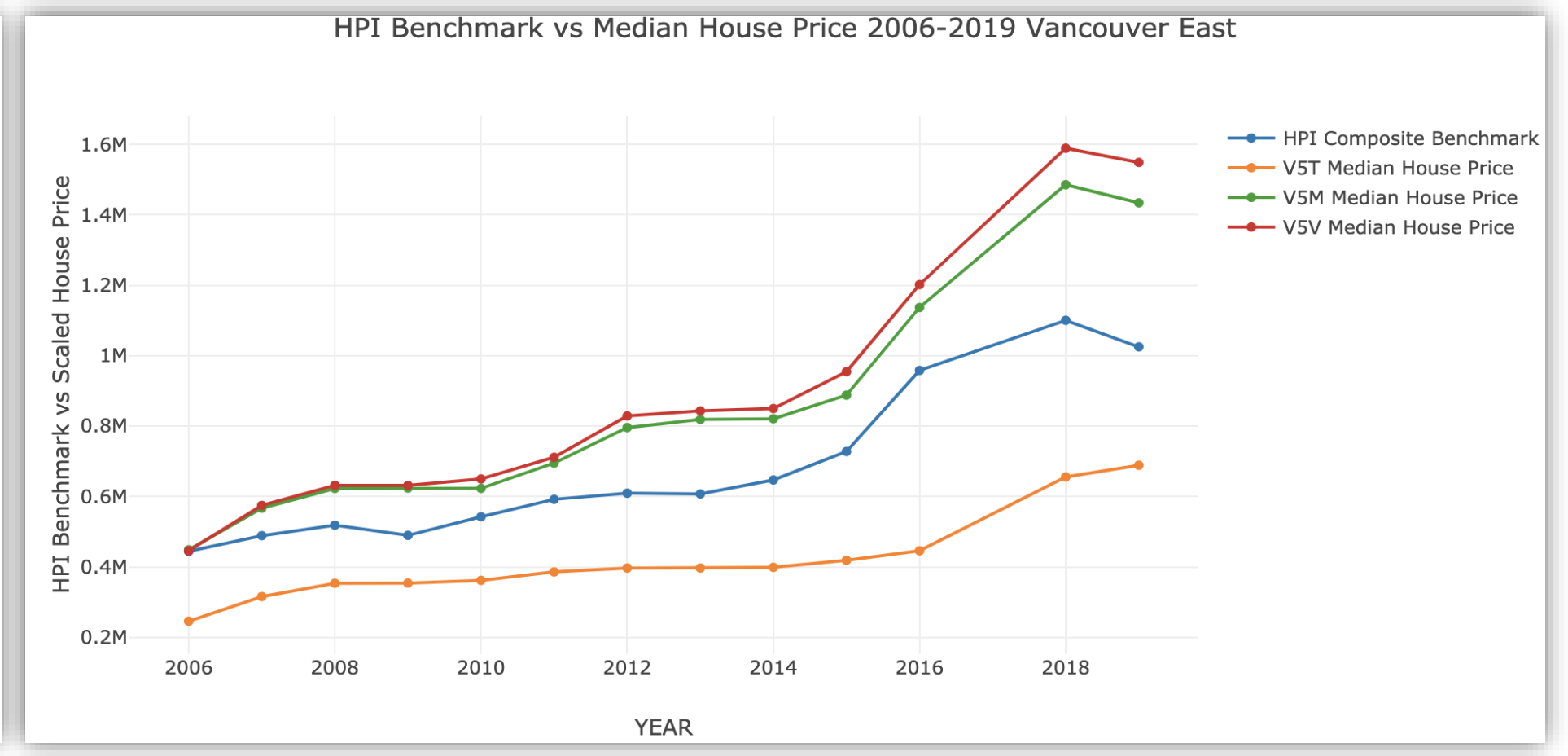
Prediction

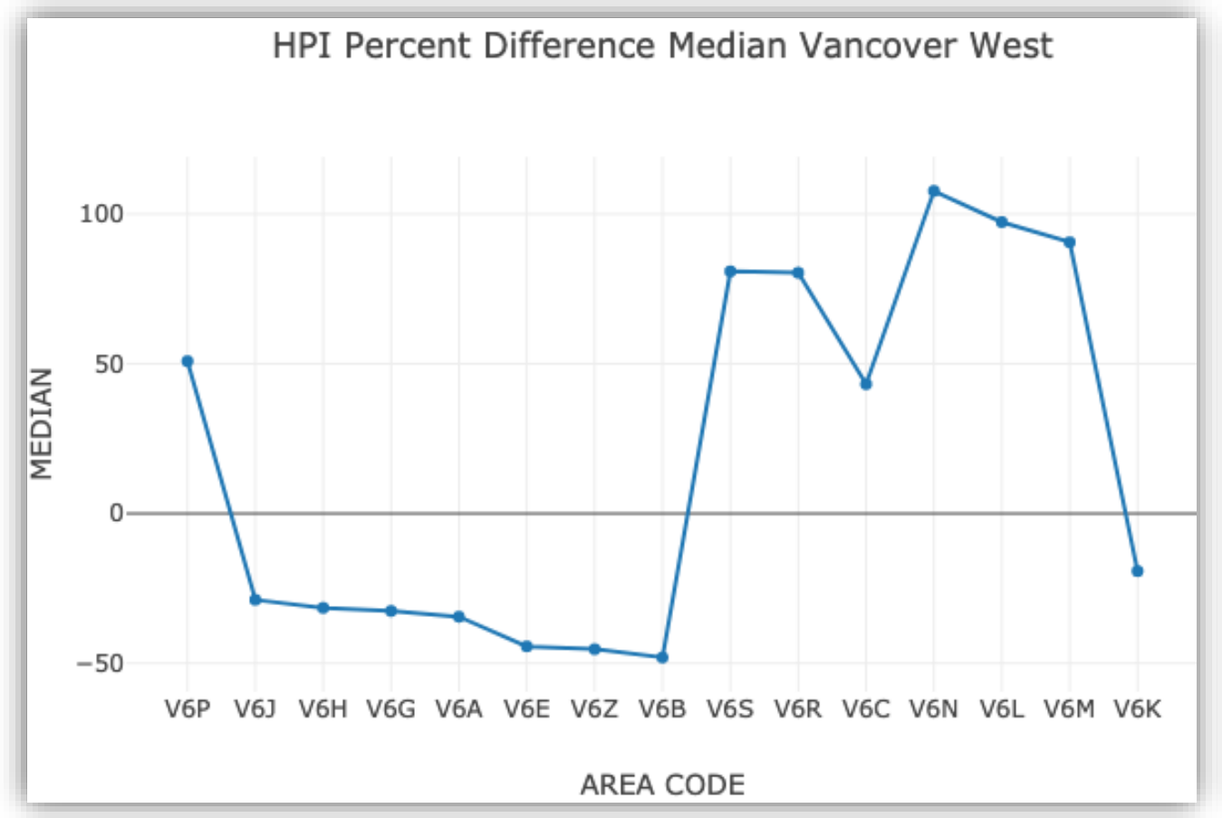
- The Regression model predicts/suggest the selling price of the property on the basis of several features using GBT regressor.
- Built a model to predict the price range based on the property image provided user.
- Property and Area wise HPI Benchmark Price future trend prediction (multivariate LSTM,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

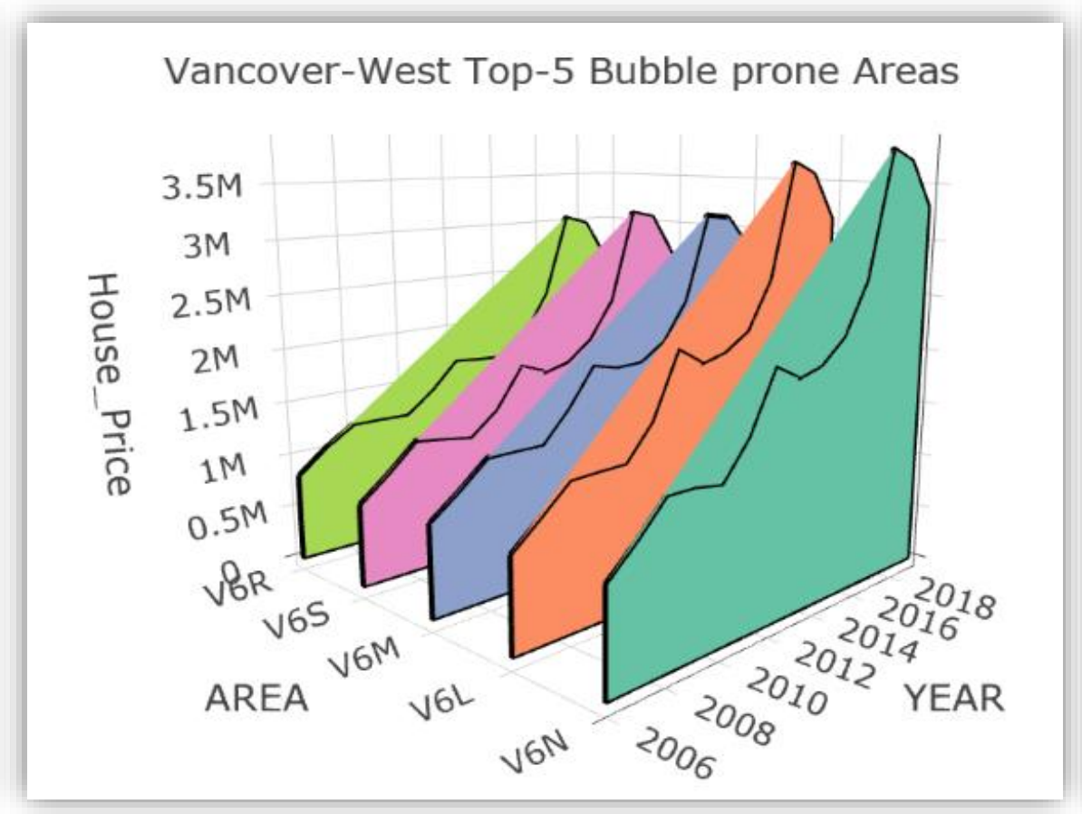
BUBBLE ANALYSIS

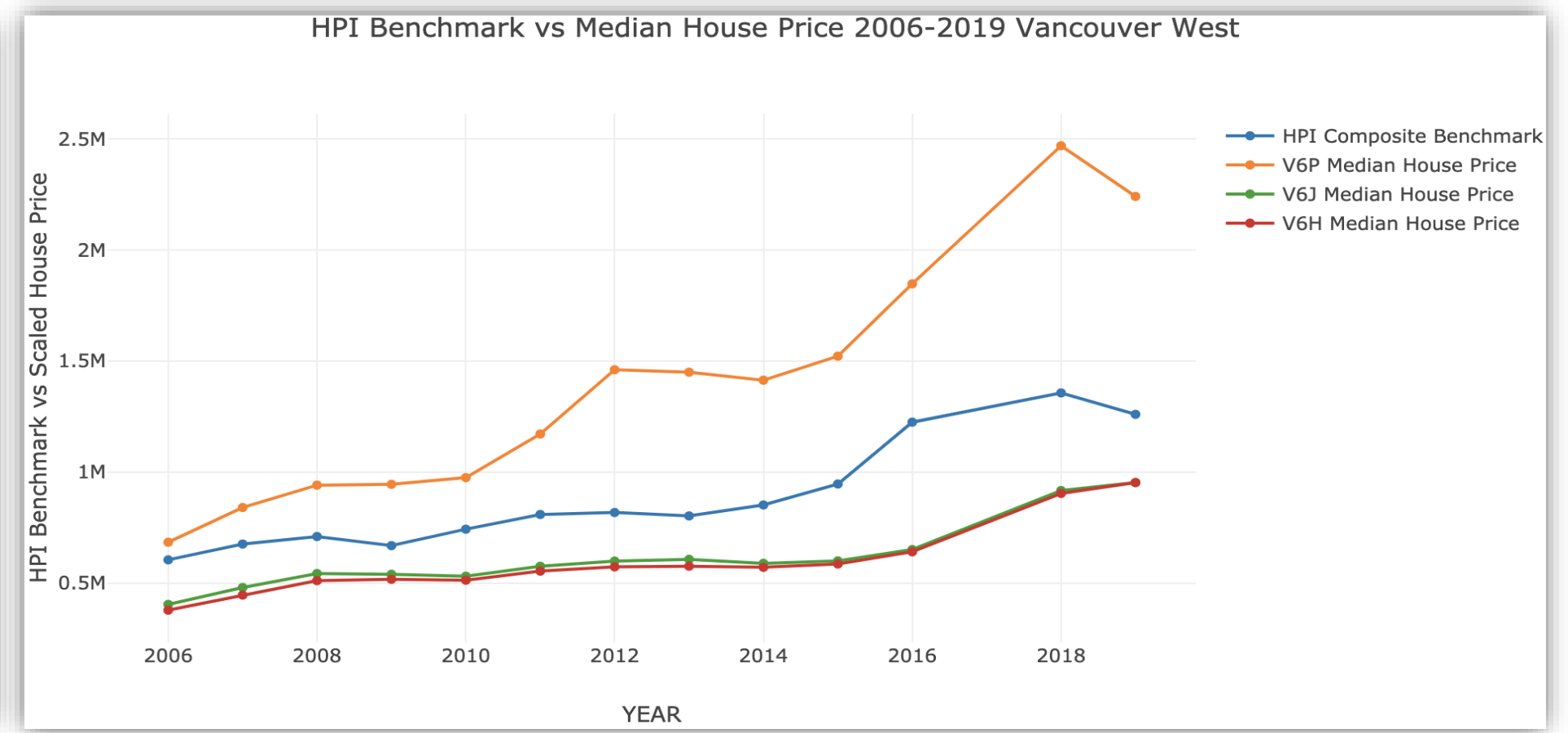






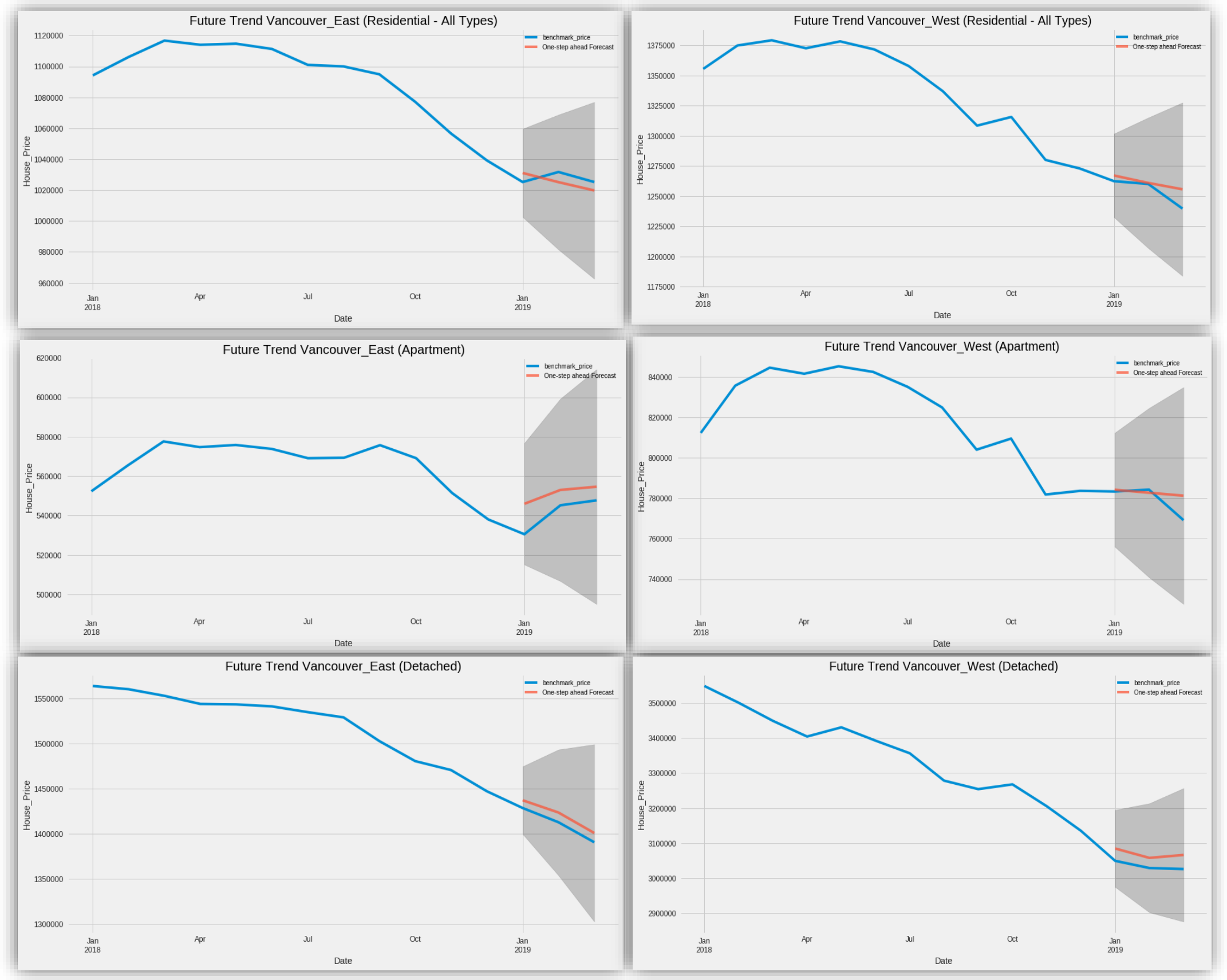






- With proper analysis of the past 15 years data(Property Tax Report, REBGV HPI Index), we have found top 5 areas are most PRONE to housing bubble. This is done through the analysis of area wise house price median percentage changes w.r.t HPI Composite Benchmark Price(HPI Index is calculated using multivariate regression analysis, a commonly used statistical technique.) from the year 2006 to 2019.
- To calculate area wise percentage changes, for all records we have calculated **Percent Change = ((HOUSE PRICE – Composite Benchmark) / Composite Benchmark) * 100**. The records which are falling outside the threshold percentage (-10 < PC > 10) are considered as vulnerable to bubble.
- From the above result, all the areas which are showing up 7 or more times(years) in the span of 15 years are marked as Highly Vulnerable areas.
- Top 5 areas obtained from the above have been Considered as our BUBBLE PRONE AREAS.
- For the top 3 bubble prone Areas from Vancouver east and west we can observe the pricing trends from the past 15 years in comparison to the HPI bench mark price. Out of 3 areas, some of them are having the pricing way above the HPI while others are below the bench mark, which clearly states such areas are prone to uneven price fluctuation.

FUTURE HPI BENCHMARK PRICE PREDICTION



- The dataset used fro this problem is the HPI Benchmark Price provided by REBGV.
- We used LSTM and SARIMAX for HPI Benchmark Price future trend prediction using monthly REBGV HPI dataset from 2006 to 2019. We converted a time-series problem into supervised learning by using sliding window transformation.
- The plots represents HPI Benchmark Price predictions for different property types in Metro Vancouver.
- We used learning and predictions from this model to enhance bubble price prediction.
- We took into account the seasonality and trend and used Dickey-Fuller test to identify the stationery propertied of the data. We remove the seasonality by subtracting the shifted series. For property type with fewer points SARIMAX provided comparable result to our LSTM model.

LEARNING AND NEXT STEPS

- We learnt and built the whole data science pipeline for the project which has a potential to be a commercial product.
- Data Wrangling is painful but fruitful. Having dealt with multiple data sources, building a pipeline was a tough part, we had to modify our approaches multiples times because of the data inconsistencies. Having said that, after getting the data under proper shape, we were able to draw interesting and useful insights/predictions.
- Building and interactive Web Frontend, which will help users to take informed decision based on the trends currently prevailing in Metro Vancouver. Also, be able to get a suggested selling price for their property . Predicting property price range based on property image.
- Build a interface where in all three models can share their learnings and give best possible future trend prediction and suggested selling price.
- Exploring Seq2Seq model to increase the accuracy and predictions obtained using LSTM model.