SFU Simon Fraser University

CMPT 733

Vancouver Housing Market Decoder

Yuyi Zhou Yabin Guo Junbo Bao

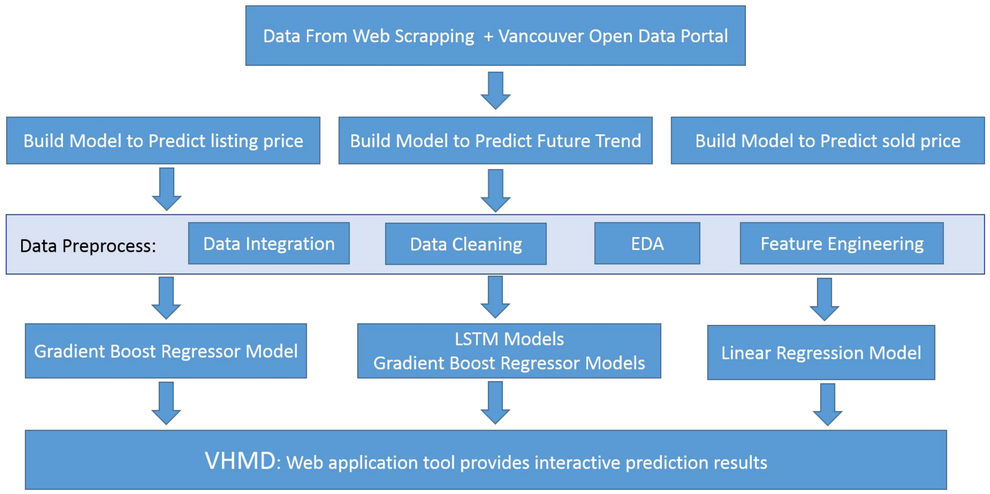
VHMD

Introduction

Most people have hard time making decisions when buying or selling a property. There always exists several questions interesting to users. What is the best selling price for sellers? What is the most possible strike price? How will these prices change over time?

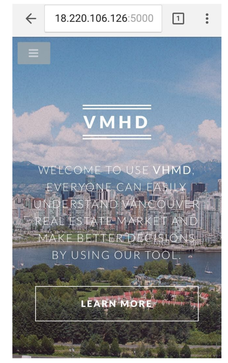
To answer these questions and help people make inform decisions, we started from collecting data and built several predictors. The predicting functions are presented in a web based app. We target to help three types of users, seller, buyer and mortgage manager. Our product is user friendly. It provides suggested listing price for sellers, property value future trends for mortgage manager and estimated sold price for buyers. Instead of checking the information from different websites, the user can find all the necessary information with our app. We combined the map and the result visualization to help user understand Vancouver real estate market more easily.

Methods & Processing Pipeline



Results & Conclusions

In this project, we created three user scenarios, seller, mortgage manager, and buyers. For each subproblem, we conducted the whole processing pipeline of collecting data, preprocessing data, EDA, and building prediction models. We spent about 80% of the time in data collection and data pre process steps. We tried different models to find the best ones we used in this project. For each model, we tried our best to achieve the high accuracy.



VHMD, an interactive web based app was built and deployed on AWS EC2 to help users easily access and interact with prediction results via a web interface. Each page not only provides the functionality of predicting, but also shows the information or trend of the neighborhood which is not provided by mainstream property listing websites.

Predict Listing Price for Seller

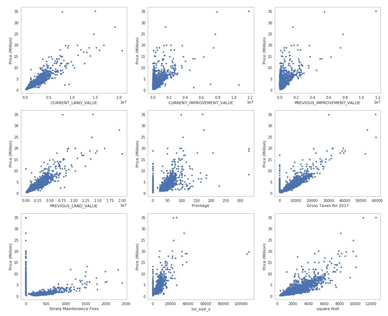


Fig. 1. Scatter plot of nine numerical features versus listing price

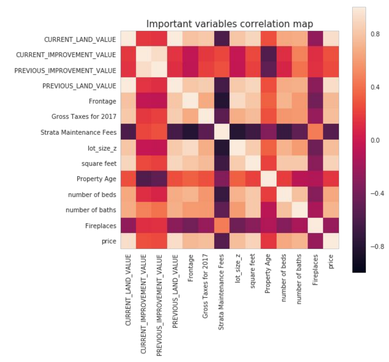


Fig. 2. Correlation between each pair of numerical variables

All the property listing information was acquired by scraping the data from rew.ca. The scraped data was combined with 2018 Vancouver tax report by entity resolution. To be specific, the Jaccard similarity between addresses from both sources was used to recognize matches. To extract the meaningful features, spearman correlations between each pair of numerical features were calculated (Fig. 2). Clearly, all nine features shown in Fig. 1 have positive relationship with listing price. In addition, each categorical variable was also plotted against price (Fig. 3). Finally, we picked twelve numerical features and six categorical features to build the predictor. The missing values are treated accordingly. The categorical feature is transformed into the vector with one hot encoding method. Eventually, approximately 2500 records collected within one month were used to build the regressor. We tried several regressors from python sklearn package on this sub-problem. The gradient boosting regressor achieved the best r^2 score (~0.964). Hence we choose this regressor as the predictor of our app.

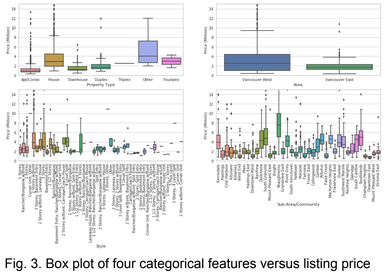
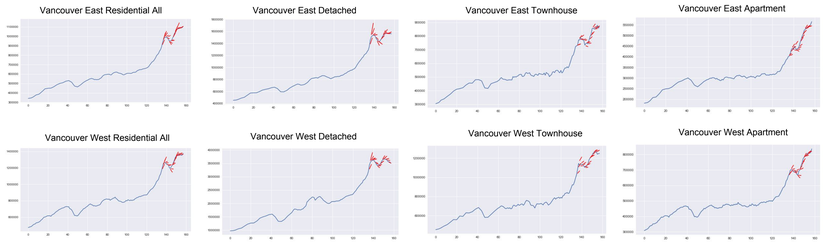


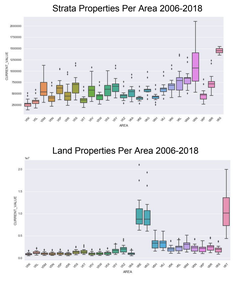
Fig. 3. Box plot of four categorical features versus listing price

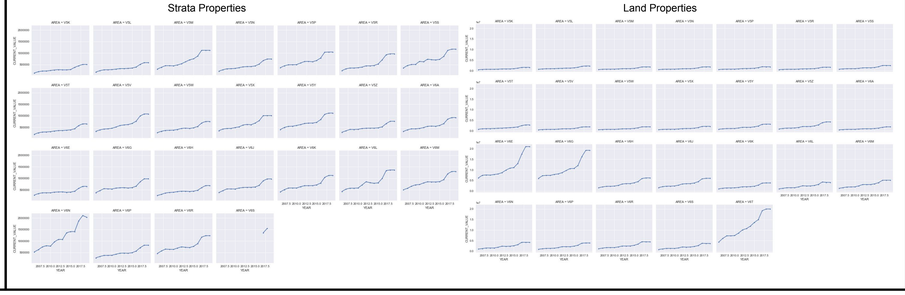
Predict Future Trend

To predict the future trend, the dataset with property benchmark value from website of Real Estate Board of Great Vancouver was used. The dataset contains benchmark property value from 2005 Jan to 2018 Feb, monthly. The LSTM model was used to predict property value from 2018 March to May. This is a multi-step time series forecasting problem. Data pre-process includes following steps: made data stationary by using the differences, transformed series into train and test sets for supervised learning, scaled the data into 0-1. There were 8 LSTM models tuned and trained for 8 types of target data. The best train results are shown below.



In addition, we also predicted the future trend for each area. The dataset was from property tax report 2006 to 2018. The average value per legal type(strata vs land) per area(first three digits of postal code) was calculated and used. The future average value was predicted to show the trend. The time series data was then transformed to fit into supervised learning model. There were 26 gradient boosting regressor models tuned and trained for land properties in 26 areas and 23 gradient boosting regressor models tuned and trained for strata property in 23 areas.





Predict Sold Price for Buyer

The data used to predict the sold price for buyers was a combination of the data from BC assessment and 2018 Vancouver tax report. We chose 14 numerical features, and from figures below we can see that they all have some relationships with the sold price. In addition, we have chose 2 categorical features house type, and the subarea, which is the first three letters of the postal code. These categorical features were transformed into the vector with one hot encoding method. Finally, to predict the sold price, we tried many regression methods, and linear regression performs the best, which achieved 0.963 of the r2 score.