VANREAPER – Vancouver Real Estate Analysis and Predictions

Chirag Ahuja, Ekramul Hoque, Pavan Kosaraju, Rohith Sooram

Table of Contents

1. Introduction

VANREAPER is an online tool which aims to improve the process of how people in Vancouver buy and sell homes, empowering them with the information they need to make a decision before making the purchase. The tool leverages the power of Data Science and Machine learning algorithms to deliver accurate information to the end-user, helping them to understand the real-estate market in a better way.

2. Motivation and Background

Buying a property is financially the biggest decision we make in life. Sellers are concerned about properly timing and getting the best price for their property. On the other hand, buyers look for demographics, getting properties as per their needs, purchasing it for a reasonable price, and know about future price trends so that they can be profitable when the property is sold in the future. Understanding the marketplace and the realities of one's personal situation is a tricky proposition and often requires the guidance of a real estate agent or extensive research. In the volatile real estate market, neither realtors' advices nor federal assessment notice is adequate for making a buy or sell decision. Given the enormity of the purchase and complexity of the factors that go into the decision, the situation easily leads to decision paralysis.

To mitigate these concerns and simplify the decisions for the end-user, we have come up with VANREAPER. Our tool guides the end-users in the right direction whether they are selling or buying the property with the help of data science and machine learning tools.

3. Problem Statement

Our data product aims to predict the property's market value to help making buy or sell decisions and determining the right timing. We provide an intuitive and easy to use web front-end encapsulating machine learning models to simplify end-user's problems. For example, the tool could monitor the advertising websites, and retrieve data from these websites to a database, then use machine learning algorithm to predict the property's current value when the attributes of the property (built year, type, Area size, Postal area, PID etc.) are specified. With the help of these predictions, our tool recommends properties which are similar to the end-user's requirements, thereby eliminating the need to visit multiple websites.

For people who are looking for places in a specific locality, our product will provide area level predictive analytics of the real estate once the user selects name of the area from the dropdown. We have leveraged the power of Google Maps to visualize both - the results of recommendation systems and area level predictive analytics provided by our backend machine learning models.

Furthermore, the key interest rate set by the Bank of Canada has significant impact on the real estate market. By analysing the correlations between the historical interest rates and property value data, our tool tries to predict the future value of the property using state of the art time series models. For example, our tool visualizes what would the trend look like for the value of a specified property in the next 12 months for a given interest rate. Based on this data the end-user can decide when would be the best time to sell (or buy) the property. It is lucid that our tool is heavily dependent on machine learning models. So, the most challenging part in this project is our models should be able to predict the results accurately. Perfecting a machine learning model accompanies lot of challenges. Some of the challenges we faced are –

- To predict the property values and future trends with high accuracy, we need lot of data which encapsulates all the information and factors affecting the price. Unfortunately, Vancouver Housing Market is vast and hence there is no single exhaustive dataset which captures all the information. So, we need to scrape a lot of data.
- Scraping always comprises of discrepancies in the data, not to mention its raw data. So, we need to spend a lot of time in cleaning and eliminating such discrepancies by deleting duplicate data, removing outliers, etcetera.
- Since our data is scraped in multiple resources, the intricate process of data integration leads to a lot of features.
- Housing Rates are dependent on numerous factors, leading to difficulties in pinpointing exactly which combination of features have the most impact on the price of the property.
- There is no perfect model in the machine learning world, so it is an exhaustive process to try out different models with all possible parameter tunings to select the best model since.
- Conveying our results and information in a simple way to the end user is perhaps the most important challenges because of the complexity of the real-estate domain.

4. Data Science Pipeline

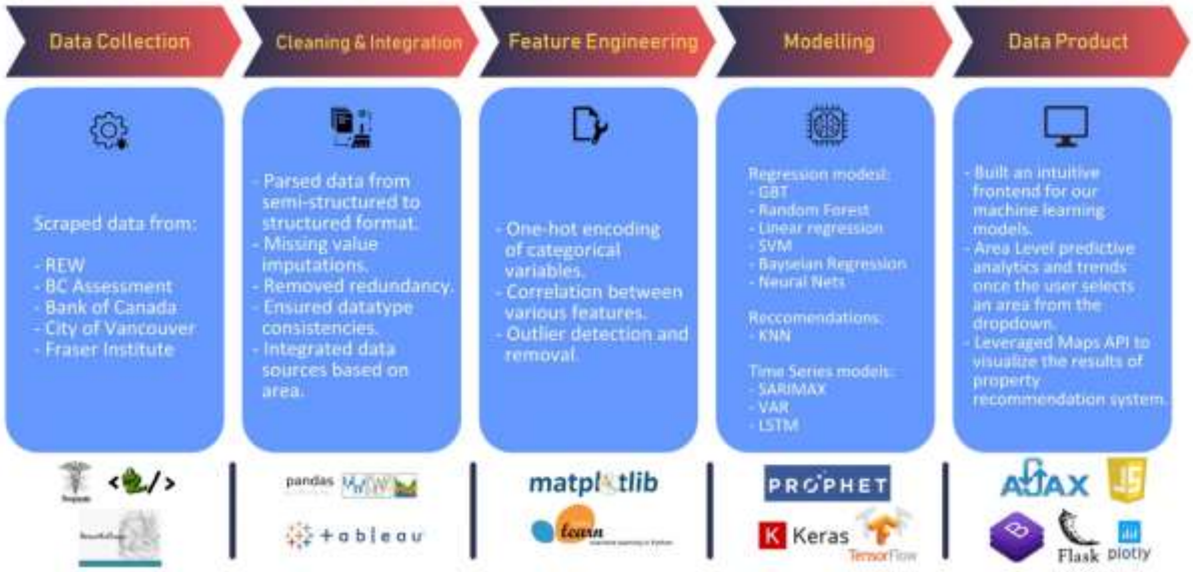

Data Collection
Cleaning & Integration
Feature Engineering
Modelling
Data Product
Scraped data from:
- REW
- BC Assessment
- Bank of Canada
- City of Vancouver
- Fraser Institute
- Parsed data from semi-structured to structured format.
- Missing value imputations.
- Removed redundancy.
- Ensured datatype consistencies.
- Integrated data sources based on area.
- One-hot encoding of categorical variables.
- Correlation between various features.
- Outlier detection and removal.
Regression modest:
- GBT
- Random Forest
- Linear regression
- SVM
- Bayseian Regression
- Neural Nets
Reccomendations:
- KNN
Time Series models:
- SARIMAX
- VAR
- LSTM
- Built an intuitive frontend for our machine learning models.
- Area Level predictive analytics and trends once the user selects an area from the dropdown.
- Leveraged Maps API to visualize the results of property recommendation system.
pandas
tableau
matplotlib
learn
PROPHET
Keras
TensorFlow
AJAX
JS
Flask
plotly

Figure 1: Data Science Pipeline

Our pipeline has a total of 5 building blocks consisting of Data Collection, Cleaning and Integration, Feature Engineering, Modelling, and the final Data Product. Initially we have gained some knowledge about our problem statement and realized that most of our data needed to be scrapped. After collecting the data from all the available sources, we need to clean our data from all the inconsistencies and the duplicate records and transform them into a structured format by storing them in CSV files. Once the data is cleaned, we need to map all the records accordingly so that we have all the factors affecting the property prices are in one place. Next, we need to analyse our data and see how all these factors are dependent on each other. This is done with the help of collinearity measures and various plots which will help us infer the dependencies of features. Additionally, there are lot of outliers in the data which hinders our linear measures, so we need to remove the outliers and clean our data once again. Also, we need to one-hot encode our categorical variables so that our models can be trained well. Once we train our models, we encapsulate them with APIs and provide an interactive front-end for the end-user.

5. Methodology

5.1 Data Collection

In order to avoid being banned by the websites from scraping their data, we gave a random pause before sending each request along with sending every request from a different IP address. This made sure that website being scraped perceived the scraping requests as organic traffic. We deployed numerous web scrapers with rotating IP addresses on the backend to scrape data from various sources which are listed below –

- REW (https://www.rew.ca/): In order to obtain recent property listings, this site has been used as a source. It provides details such as bedroom, bathroom, area square feet, age, type i.e. townhouse or apt/condo that has been used for our recommendation model. After filtering and removing outliers, the dataset contained approximately 12K records.
- BC Assessment (https://www.bcassessment.ca/): The dataset provides all the property listings in Vancouver. Along with the property tax report the dataset contains postal codes, current improvement value, previous improvement value etc. This dataset allowed us to sort prices vs year and postal code for each listing id. It contained more than 200k records.
- Bank of Canada (https://www.bankofcanada.ca/): Part of our prediction hypothesis and time series analysis involves interest rate as a strong influencer. The dataset provides interest rate from year 2006 – 2019.
- Fraser Institute (https://www.fraserinstitute.org/): The site provides school ratings available for each area. In our EDA we analysed how school ratings and its distance are correlated with a property price.
- City of Vancouver (https://vancouver.ca/): Crime data for each area has been collected from this site. This has been used to understand how price trends behave with change in crime rate in a specific area.

Following this scraping process, the collected real estate data accounted to more than 1GB of raw and uncleaned data.

5.2 Cleaning and Integration

The data collected was semi-structured in nature and hence had to be converted into tabular format. We used parsing tools like 'lxml' and 'BeautifulSoup' to convert HTML files into CSV files. Before directly using the data in the csv files to perform analysis, we had to clean the data to remove the irregularities in the data.

We removed the outliers in the data by using the median absolute deviation (MAD) methodology since it is robust to outliers. Anything that more than 3 standard deviations away from the median was treated as an outlier and discarded. Average absolute deviation from the mean was not considered as this method is susceptible if the data contains many outliers.

The property listings data from REW and rental listings from Craigslist both have a unique identifier which identifies each property. This means to remove duplicate records we simply had to find out the if the listing id has been repeated in the entire dataset. The records which had missing values were imputed with median substitution method instead of deleting the records because deleting records with missing values may lead to decrease in the power of analysis by decreasing the effective sample size.

In addition, we performed numerous aggregations and joins to transform the data into a format that is fit for further analysis. The tax data from the years 2006 and 2019 was joined and aggregated to get the required median property prices for each region in Vancouver and their corresponding interest rate in those years. This data was further used for time series analysis and predictions. Data was aggregated and integrated to get statistical results such as median price of the property, age of the property, property type distributions, schools and their ratings for different areas in Vancouver. The rental data was cleaned and parsed to extract fields like number of beds, location, size of the property in square feet and the rent.

5.3 Exploratory Data Analysis & Feature Engineering

Most of the features we collected like postal code and house type were categorical in nature. We converted these categories into real values using one hot encoding as machine learning models insist on predicting a numerical value.

We found that when we categorize the school ratings data into 3 categories, we observe a strong positive correlation between school ratings and the property prices. The first category consists of schools which are rated between less than 6, the second category of schools rated between 6 and 8 while the third category had schools rated more than 8. This correlation has been visualized in Figure 2.

We investigated whether there is a correlation between school distance and property prices but there seems to be no correlation between them. People care more about the quality of the school than the distance. This finding has been depicted in Figure 3. It seems that people care more about the quality of the school than the distance.

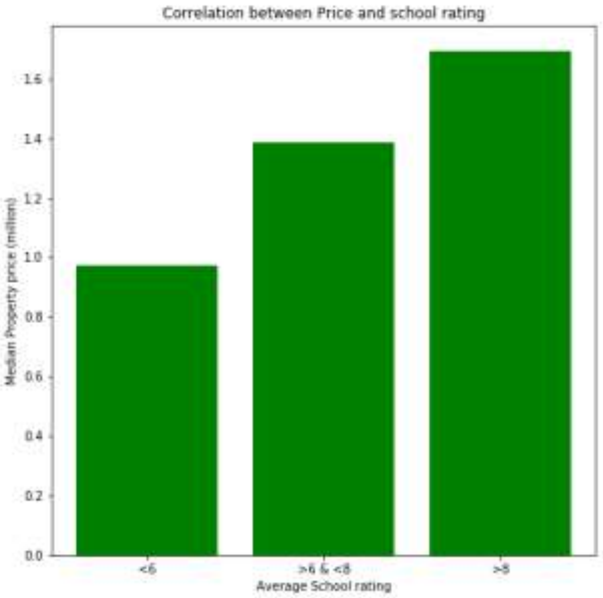


Figure 2: Correlation between school rating and house prices

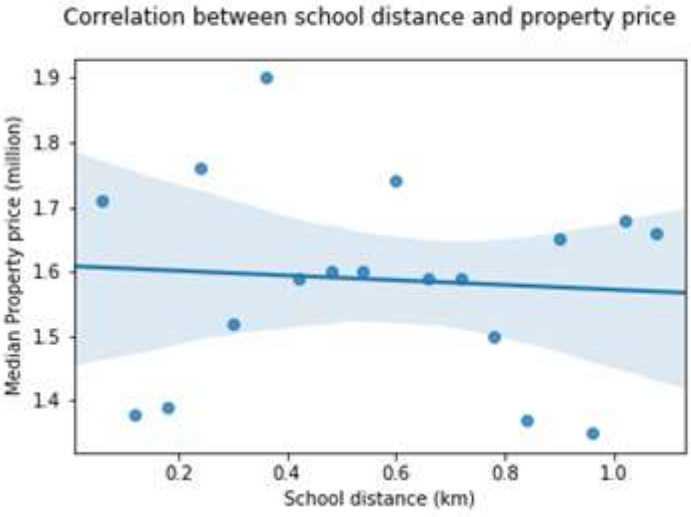


Figure 3: Correlation between school distance and property price

Another hypothesis we had was that the property rates must also be must be sensitive to the crime in the surrounding areas. We confirmed our hypothesis by plotting the frequency of crime vs property rates. The plot in Figure 4 confirms our intuition that there does indeed exist an inverse correlation between number of crimes in a neighbourhood and median property price in an area.

The next relationship we wanted to analyse was whether the age of the property has an effect on its price. We uncovered that the older properties are fewer in number but are worth more than newly built properties as depicted in Figure 5. The reason for this trend is that older properties have been built using better quality construction materials and generally occupy a large amount of area in terms of square feet.

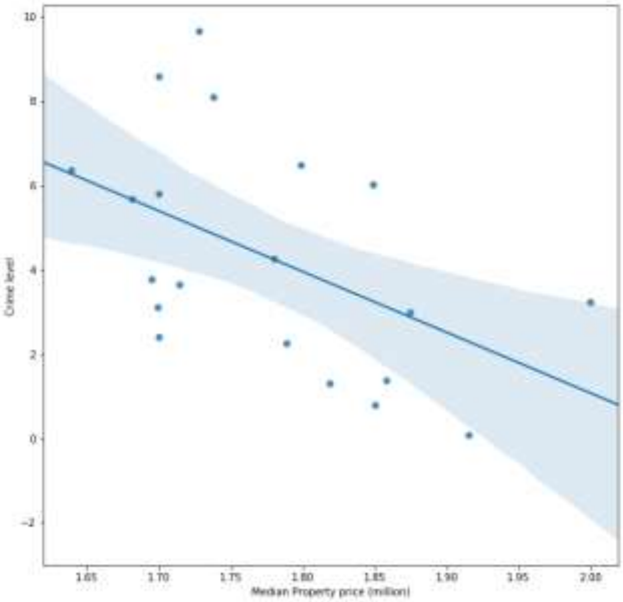


Figure 4: Inverse correlation between crime level on y axis and property price on X-axis

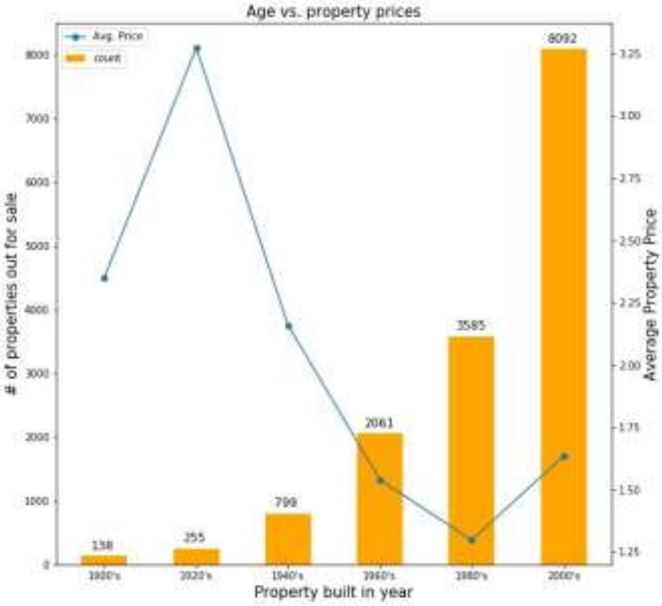


Figure 5: Plot showing how older properties are fewer in number but are more expensive than newer ones

We also plotted historical interest rates in Canada versus the median property values prevailing in Vancouver and discovered that property prices and interest rates are inversely correlated. This makes sense because as the interest rates become cheaper, the loans get cheaper and people want to take the risk of owning a property while paying lower interest rates. This drives the demand for housing higher and thus the prices go up. The inverse correlation between interest rates and property prices has been visualized in Figure 6.

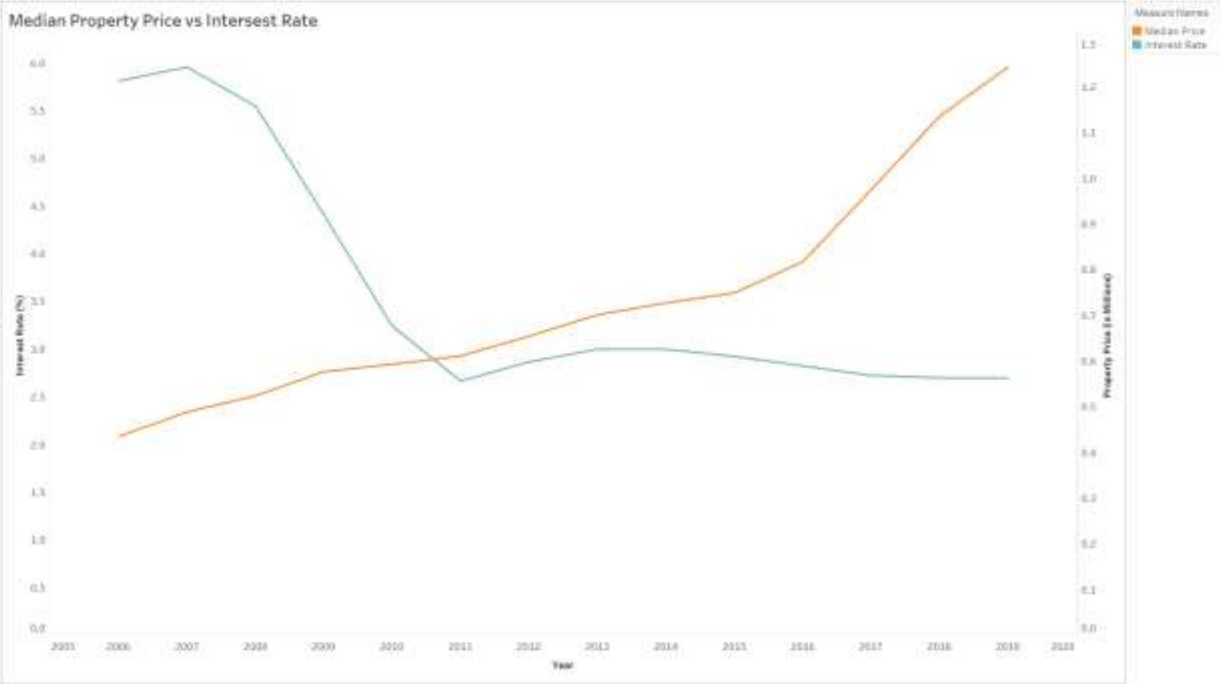


Figure 6: Inverse correlation between price and house prices

The conclusion of our exploratory data analysis is that the features that most affect the housing property rates are area measurements of the property, number of bedrooms and bathrooms, age of the property, crime rate in the area, school ratings and prevailing interest rates. The correlation matrix heatmap of all the features affecting the property price has been depicted in the Figure 7.

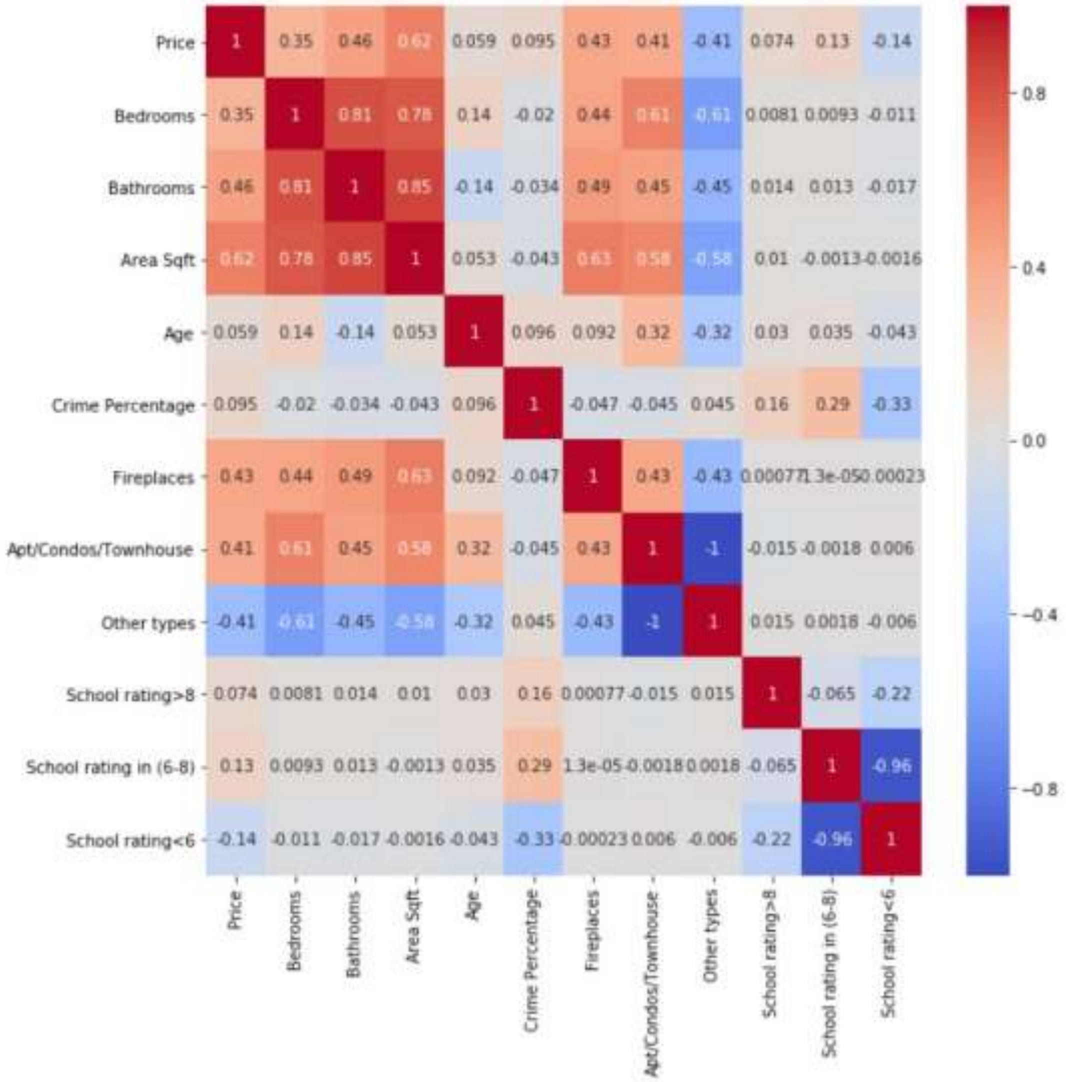


Figure 7: Correlation Matrix

5.4 Modelling

Our project incorporates regression model, time series model, and a simplified recommendation engine. A lot of optimization and testing methodologies went into the process of choosing the best model for each segment of our project.

Regression Models: To predict the housing prices, we started off by trying regression models like Gradient Boosting Trees (GBT), Random Forest Regressor, Support Vector Regressor and Neural Networks. We did hyperparameter tuning for all the above models but the results were not encouraging because of the paucity of data and curse of dimensionality. To creep our accuracy scores up, combining different predictive models seemed a better approach. So, we stacked the models and applied a linear regression model on top of it to obtain the property prediction. This has improved the prediction scores further. We integrated this stacked regression model to our product which will predict the house prices when the end-user enters a certain set of features.

Recommendation Models: Based on the property price predictions, our product also recommends some properties which are similar to the user-entered features. To provide accurate recommendations to buyers looking for properties, we utilized KNN (K Nearest Neighbours) algorithm to come up with property suggestions for the end user. KNN is a non-parametric, lazy learning method. It uses a database in which the data points are separated into several clusters to make inference for new samples.

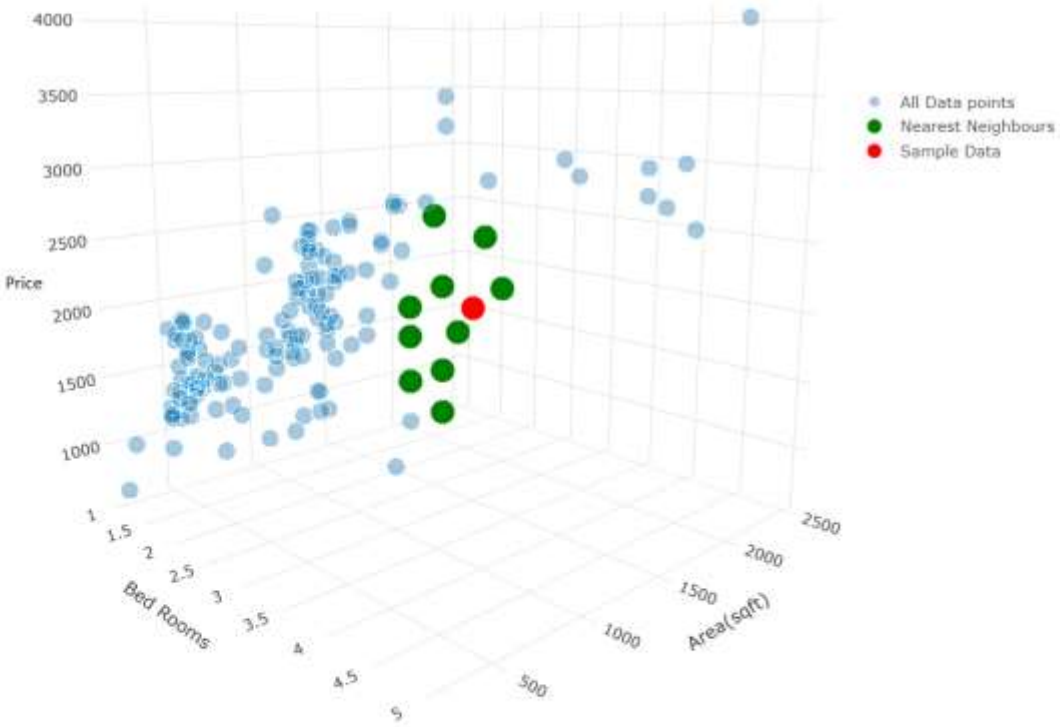


Figure 8: 3D plot of 10 nearest neighbours in the Recommendation System

KNN does not make any assumptions on the underlying data distribution but it relies on item feature similarity. When KNN makes inference about a property, KNN will calculate the "distance" between the target property and every other property in its database, then it ranks its distances and returns the top K - Nearest Neighbours properties as the most similar property recommendations. For property recommendations we chose value of K as 30 thereby recommending 30 properties similar to the user's inputs. The plot for K nearest neighbours, for K as 10 has been depicted in Figure 8.

Time Series Models: To predict the future trends of an area and property, we have integrated time series model into our data product. To get the best results, we have tried various models like ARIMA, VAR, LSTM, and chose the one which outputs the optimal results.

Autoregressive Integrated Moving Average (ARIMA) is popular and widely used statistical model used for time series forecasting. It is a class of model that captures a suite of different standard temporal structures in time series data. The key aspects of ARIMA model are:

- Autoregression: A kind of model that uses the dependent relationship between an observation and some number of lagged observations
- Integrated: The use of differencing of raw observations in order to make time series stationary
- Moving Average: A model that uses dependency between an observation and a residual error from a moving average applied to lagged observations.

The parameters of ARIMA model area:

- **p:** The number of lag observations included in the model, also called the lag order
- **d:** The number of times that the raw observations are differenced, also called the degree of differencing
- **q:** The size of the moving average window, also called the order of moving average

After tuning the hyperparameters, we found the ideal values of (p, d, q) in our case was (1,0,2). Although we performed hyperparameter tuning, the performance of the model on the validation dataset was disappointing.

The next time series statistical model we tried was Vector Auto Regression (VAR) model. The Vector Autoregression model generalizes the univariate ARIMA by allowing for more than 1 independent variable. In VAR, each variable has an equation has an equation explaining its evolution based on its own lagged values, the lagged values of other model variables and error term. So that means the property prices in 2019 are dependent on the property prices in 2018, property rates in 2017, property rates in 2016 etc. as well as on the interest rates in 2018, interest rates in 2017, interest rates in 2016 etc. The parameters of the model are fitted using the maximum likelihood equation. For our case, the VAR model performs much better than ARIMA.

Prophet is a time series forecasting library implemented using R by Facebook. The prophet library is completely plug and play in usage and provides completely automated forecasts. Under the hood, Prophet uses additive regression models with 3 main components namely yearly, weekly and daily seasonality plus holiday effects. In our experiments with Prophet, it came out ahead of both ARIMA and VAR models for time series prediction of property rates on the validation set.

Next up, we tried our hands at deep learning methods like LSTM neural networks for time series forecasting. LSTM models are able to seamlessly model problems with multiple input variables which leads itself well to multivariate time series forecasting problems. Before feeding the features to the LSTM network, all the features were normalized and dataset was converted into a supervised learning format using the sliding window technique. We used a sliding window of 3 lag order as it gave us optimum results.

5.5 Data Product

We used Flask web framework to host our machine learning models. With the help of Flask, we wrapped our models into RESTful API services, enabling the ease of communication between the frontend and backend. The frontend uses JavaScript, Ajax, HTML and Bootstrap for CSS.

To develop an API for our models, we have serialized our models into pickle files. When the user enters information on the front end, a REST call is made to the backend API of the model and the results are transferred back to the frontend. All the static data which is required for the front-end plots are stored as flat files.

The plots on the front end are generated dynamically on the fly. Based on the inputs and model outputs, we have used Plotly to generate plots. We have also used Google Maps API to visualize the recommended properties so that user can know the location of recommendations. We plan to deploy our product on AWS and use S3 services in future instead of using simple flat files. More information about the data product can be found in Section 7.

6. Evaluation

We are pretty confident about the results we have obtained from our models. Among regression models, the stacked models comprising of Gradient Boosting Trees, Neural Networks and Support Vector Regression gave us the best results in terms of Root Mean Square Error (RMSE) on the validation dataset. In time series models, Prophet library gave us the best results in terms of RMSE. The RMSE values of all the regression models can be found in Table 1 while RMSE values of time series models can be found in Table 2.

Table 1: RMSE values of Regression models

Regression Models	RMSE
Support Vector Regression	0.48 Million
Gradient Boosting Regression	0.39 Million
Random Forrest	0.44 Million
Neural Network with 2 hidden layers (200-100-30 structure)	0.43 Million
Stacked Model (Linear combination of GBT, MLP and Random Forrest)	0.31 Million

Table 2: RMSE values of Time Series models

Time Series Models	RMSE
ARIMA	0.73 Million
Vector Auto Regression	0.23 Million
Facebook Prophet Library	0.22 Million
LSTM model	0.67 Million

Even our plots suggest that our results look promising. For our recommendation system, as we can see from Figure 8, the model is recommending the properties which are closest to the input features. Our time series model also predict the trend correctly because we have tested our model and validation set and our model estimated the trend accurately. In Figure 9, the orange line is the estimated trend by our model and the blue line is the actual trend. Clearly, our model outputs good results.

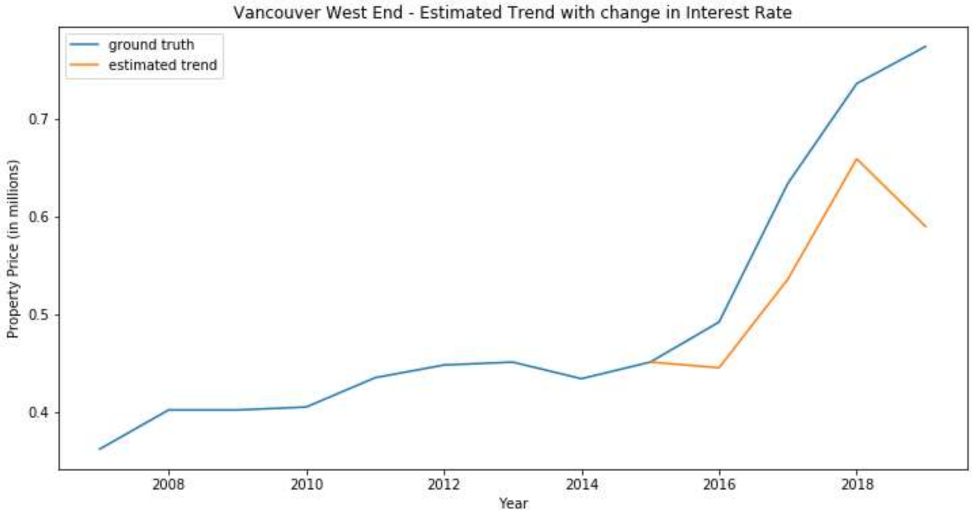
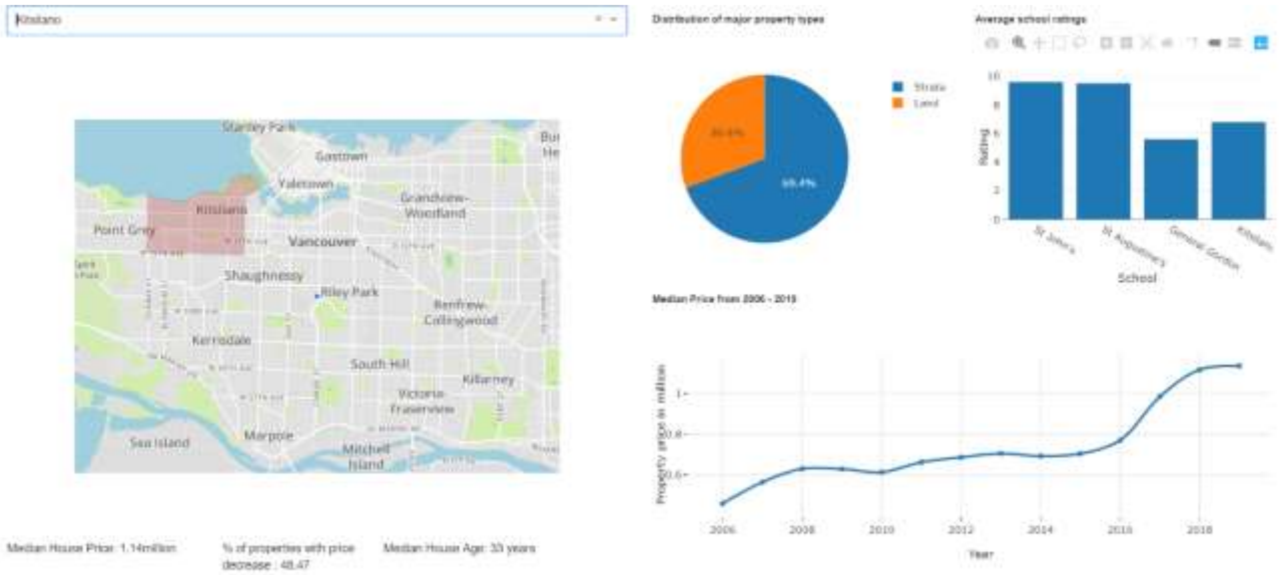


Figure 9: Time series model predicting the future trend

7. Data Product

In order to facilitate the user to use our predictive and recommendation model, we built an interactive web application. It has three views, one page covering the property price prediction and recommending the property based on user's preferences, one page to exhibit the basic analysis like historical prices and school ratings of an area, and the final page for showcasing the future trends of the property entered by the user based on its historical data.

Area analysis: In this page, user can choose an area from the dropdown list to get the basic statistics and analysis of that area. We show the historical property prices of that area, number of schools and their ratings in that area, Strata and Land types distributions in that area, and average house prices in that area. This page allows users to get an overview picture of a particular area.



Kitsilano

Median House Price: 1.14million

% of properties with price decrease : 48.47

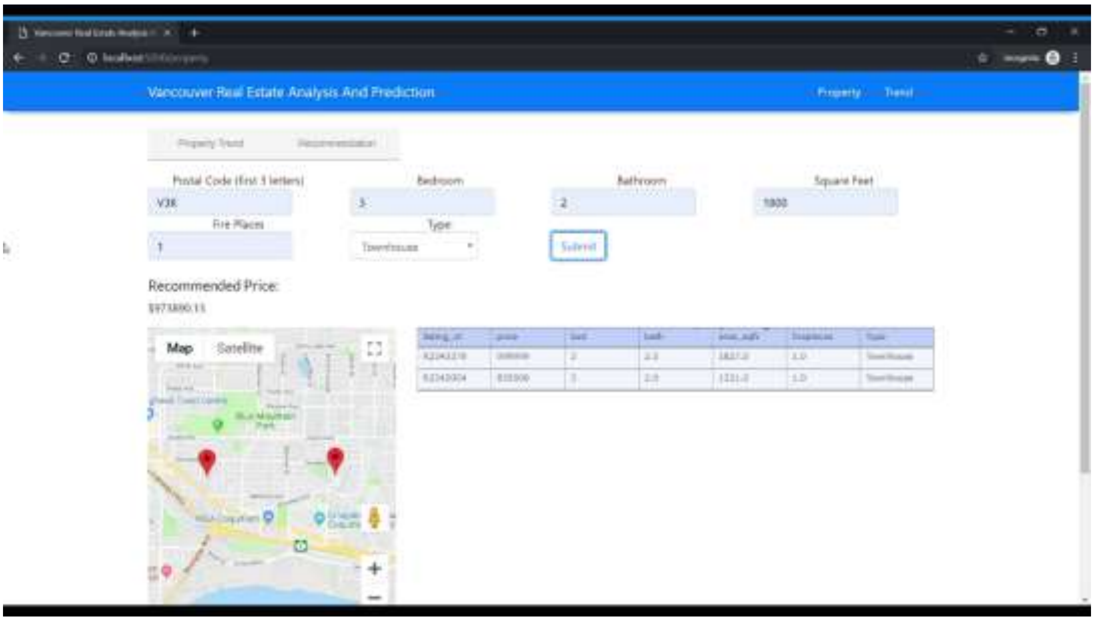
Median House Age: 33 years

Figure 10: Area analysis Page

Prediction & Recommendation: This section is particularly aimed at property buyers who are struggling to decide on a budget. The property prediction and recommendation section take the following inputs from the user:

- Number of Bedrooms
- Bathrooms
- Area Square feet
- Fireplaces
- Property Type: Apt/Condo, Townhouse etc.
- Postal code

Based on these inputs, our stacked regression model predicts the estimated price of the property. With estimated price and the input features, our recommendation model suggests the properties available at present which are similar to the user's input. These recommendations are also pinpointed on the map using Google Map APIs, eliminating the need for user to search for the location. With this module, people who are looking to buy a property will gain some insights on estimations of their budget and if their budget meets our recommendations, they can proceed with the purchase of the property.



Vancouver Real Estate Analysis And Prediction

Property Trend | Recommendation

Postal Code (first 3 letters): V3K

Bedroom: 3

Bathroom: 2

Square Feet: 1800

Fire Places: 1

Type: Townhouse

Recommended Price:

$973890.11

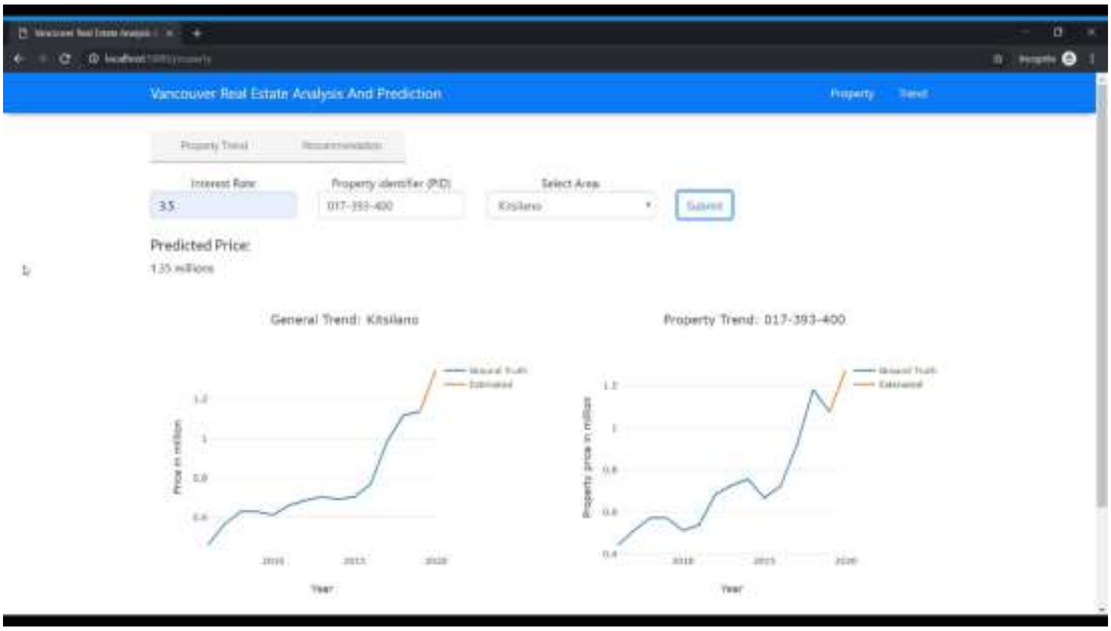
listing_id	price	bed	bath	area_sqft	fireplaces	type
R2242218	[illegible]	2	2.5	1827.0	1.0	TownHouse
R2242004	835500	3	2.5	1221.0	1.0	Townhouse

Figure 11: Prediction and recommendation page

Property & Area Trend: For property sellers who are concerned about when to sell their property, our property and area trend section can help them take the right decision. Here, initially user enters the following three inputs:

- Interest Rate
- Property Identifier (PID)
- Area

Upon submitting the information, our well-trained time series models will predict how the trend is going to look like for the next 12 months for that area and the property. This trend is visualized in two plots, one visualizing the trend of the area, and the other depicting the trend of that PID. The plots also exhibit the historical prices of that property and area for a better understanding. This will help the seller to decide whether he should wait or sell his property immediately to get the maximum profit.



Vancouver Real Estate Analysis And Prediction

Property Trend | Recommendation

Interest Rate: 3.5

Property identifier (PID): 017-393-400

Select Area: Kitsilano

Predicted Price:

1.35 millions

Figure 12: Property & Area trend Page

8. Lessons Learnt

The project taught us some important lessons about the real-world data science process

- *The importance of collecting more data*: Sometimes in the real world, improving model accuracy is more about collecting more labelled data than applying advanced algorithms on limited data.
- *Feature engineering*: Real world data is messy and the scraped data did not have the features in the form that can directly be ingested by the models. We had to perform feature engineering to transform the features into a format that can be fed into the models.

9. Summary

We have scraped data from multiple sources like rental data from craigslist, property tax data from BC Assessment Authority, property listings data from REW, school ratings data from Fraser Institute and historical interest rates from Bank of Canada. In the next step, we performed exploratory data analysis and feature engineering to get the data in proper shape for feeding it into predictive models. Regressions models like Gradient Boosting Trees, Random Forrest, Linear Regression, Ada Boost and Time Series models like ARIMA, LSTMs and Vector Auto Regressor were used to predict future prices of the properties. Property Recommendation System was implemented using K-Nearest Neighbours algorithm in Scikit-learn. Finally, all the statistical models were serialized, persisted and deployed using an interactive Flask Web Application.

10. References

1. Property value prediction with market data, Marco Wu.
2. Vancouver Housing Market Decoder, Yuyi Zhou,Yabin Guo, Junbo Bao.
3. Why do stacked ensemble models win data science competitions?, Funda Güneş, source here.