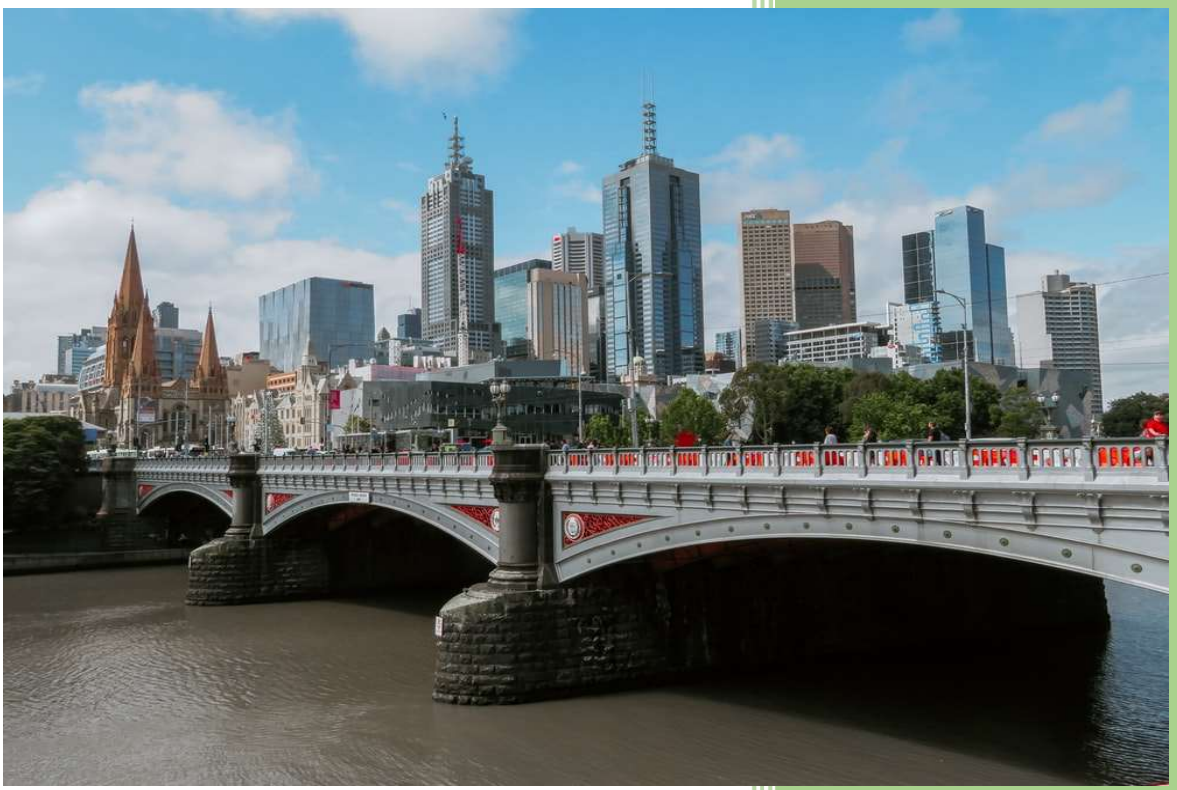
2020

IBM Applied Data Science Capstone Project Report

ESTIMATION OF HOUSING PRICES OF 3 TYPES IN MELBOURNE UTILIZING REAL ESTATE AND NEAR BY VENUES DATA



Lajpat Rai

9/7/2020

Contents

Introduction

Background

Melbourne is the capital and most populous city of the Australian state of Victoria. Today, it is a leading financial center in the Asia-Pacific region and ranks 15th in the Global Financial Centers Index with enormous employment and business opportunities. These recompenses have gained attention of several local Australians and international immigrants, aspiring them to settle and lead a good lifestyle in the city of Melbourne. But with all these amenities, there lies Melbourne's real estate market hype. For the last few years, significant influx of locals from other provinces of Australia and immigrants from worldwide have outpaced housing demand in comparison to supply. This has lead prices inch steadily upward developing a bubble in Melbourne real estate market.

Business Problem

Immigrants and locals moving into Melbourne city founds selection of house challenging for them and their families due to hot housing market. Affordability, accommodation features and nearby facilities always remain qualifying parameters for selection of an appropriate place to live for them.

Price variation with respect to housing attributes for different Melbourne neighborhoods in conjunction with nearby venues needs to be addressed to resolve the encounters faced by any individual during selection of a place. The project seeks to explore real estate data to get an insight of property price variation in combination with its traits and near services available based on location data along all the neighborhood's to establish relationship between them, which can be used as a recommendation while selecting a place to live. In addition to above, another task is to identify the potential for building more houses or apartments and where would sellers be selling off their properties, I order to maximize the profit.

Audience & Stakeholders

The target audience for this report are:

- Locals planning to settle from other provinces of Australia to Melbourne city
- Immigrants across the globe planning to settle in the city of Melbourne
- Potential house sellers and buyers residing in Melbourne city who can optimize their advertisements & profits.
- Potential business investors setting up their businesses in different regions of Melbourne city and aims to accommodate their workforce
- City planning authorities to set up more amenities in neighborhoods with less venues

Data

This section describes the data sourced for this project.

Data Sources

This project integrates data sources such as kaggle, Google Map, as well as Foursquare data. This section describes each of these data sources and provides examples of the data.

Kaggle

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. The website contains different database of various real estates. In view to the scope of the project, family living options such as unit, Condo & Town Houses in 2 regions (Southern & Metropolitan) in Melbourne city are considered. This data aims to provide details such as property type (unit/Condo/Town House), price, details such as number of bedrooms, bathrooms, car spot with their respective addresses. Since the data is available on kaggle website, the data have been downloaded in csv format from the following link;

https://www.kaggle.com/anthonypino/melbourne-housing-market

Foursquare Data

Foursquare provides a mobile app that allows users to search for near-by venues and see information and reviews. Users also feed information back to Foursquare both passively, as the app tracks users' locations, and actively as users enter venue names, locations, and reviews.

Since 2009, users have provided Foursquare with location data on over 105 million venues, with over 75 million tips from local experts. As one of the largest sources of location-based venue data, the company describes itself as a technology company that uses location intelligence to build meaningful consumer experiences and business solutions.

This project will access Foursquare venue data for all neighbourhoods. The Foursquare venue data will particularly seek to identify venues that have significant impact on property prices. These data will then be used for subsequent comparison and categorization to provide insight to the business problem.

The Foursquare venue data are accessible via application programming interface (API). A free developer account is used to access the data from https://developer.foursquare.com/places-api.

Foursquare Data Sample

Following is a sample of the imported data showing particularly the venues (by name) and the respective venue categories for each neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbotsford	-37.80385	144.996577	Three Bags Full	-37.807318	144.996603	Café
1	Abbotsford	-37.80385	144.996577	The Kitchen at Weylandts	-37.805311	144.997345	Café
2	Abbotsford	-37.80385	144.996577	The Park Hotel	-37.802769	144.997029	Pub
3	Abbotsford	-37.80385	144.996577	Stomping Ground Brewery & Beer Hall	-37.804683	144.991171	Brewery
4	Abbotsford	-37.80385	144.996577	Laird Hotel	-37.805309	144.993124	Gay Bar

Data Gathering & Cleaning

Data acquired after web scraping utilizing Beautiful Soup library had several inconsistencies which were resolved in following mentioned steps;

- Object type columns were changed to categorical variables for plotting purpose.
- Postal code was numerical variable hence it was changed to categorical variable.
- Duplicate columns were dropped e.g Bedroom2.
- Missing data in all columns was identified. Column (BuildingArea), which has most missing values was dropped and rows of price column was also dropped.
- False year-built columns values were removed (e.g Melbourne was founded in Aug 1835).

- Property Price column contained some outliers was removed from data frame to develop a consistent dataset.
- Some remaining missing values in columns (car & bathroom) were handled by replacing with most common type of its respective value. In Landsize column, missing values were replaced with average value.

Foursquare venues for Melbourne neighborhoods

The Foursquare application programming interface (API) was accessed to obtain the venues for neighbourhoods in data frame df. This project also aims to understand any relation between neighbourhood nearby venues with respect to average property price. In this regard a data frame was grouped based on neighbourhood with average property price for each type (house, unit & townhouse) for each region (Southern & Northern Metropolitan) was developed

After developing data frame, venue data using Foursquare API was acquired as follow;

(i) Client ID, Client secret & version was entered in order to access foursquare venue data

(ii) A function was developed which accessed neighbourhood name, its latitude & longitude from data frame and extracted venues based on that.

(iii) Search radius was defined as 500 m from respective neighbourhood latitude & longitude values with a defined limit of 100 venues per neighbourhood in each metropolitan region for each type of property.

(iv) Extracted venues data with neighbourhood name, latitude & longitude was appended to a new data frame.

(v) Northern & southern_metropolitan_venues dataframe contained features including Neighbourhood, Neighbourhood Latitude, Neighbour Longitude, Venue, Venue Latitude, Venue Longitude & Venue Category.

Methodology

This section describes the data exploration, inferences and machine learning approach that was applied and how they relate to the original business problem of gaining data insights specifically to identify property price variation in combination with its attributes and near services available based on location data along all the neighborhoods and develop relation between them.

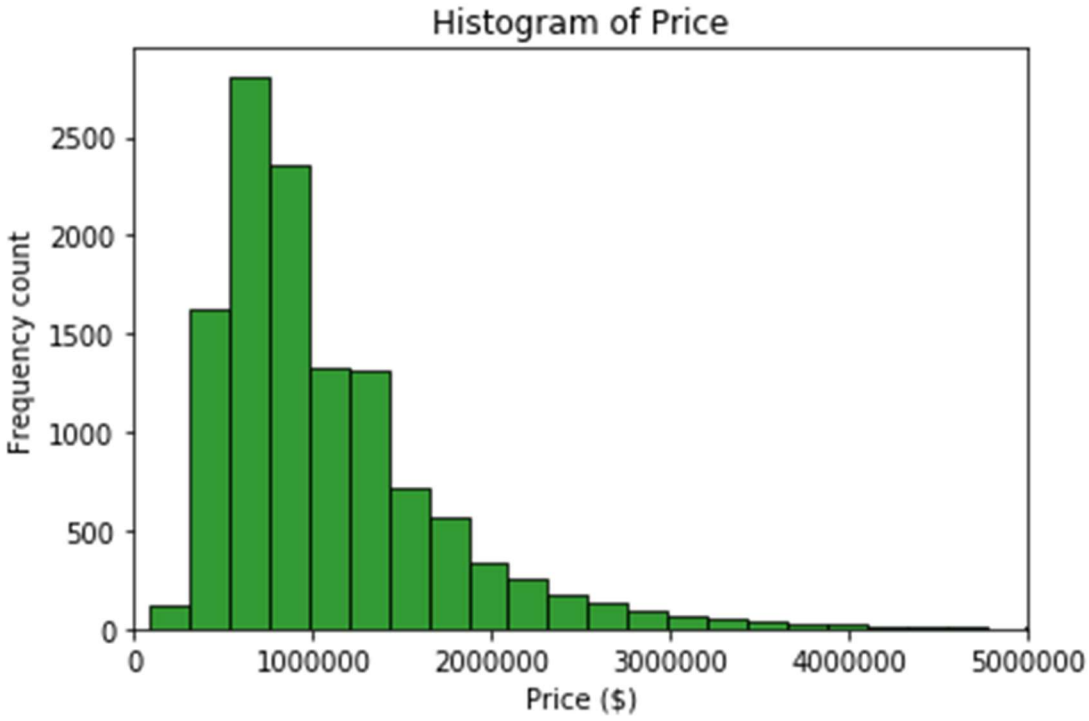
The methodology includes exploratory data analysis with the aid of histogram, boxplots, data visualization, regression analysis to investigate the influence of property attributes & nearby venues on property prices, as well as the choices and considerations within the methods.

Exploratory Data Analysis

Exploratory data analysis was conducted utilizing histogram, boxplots & barplot.

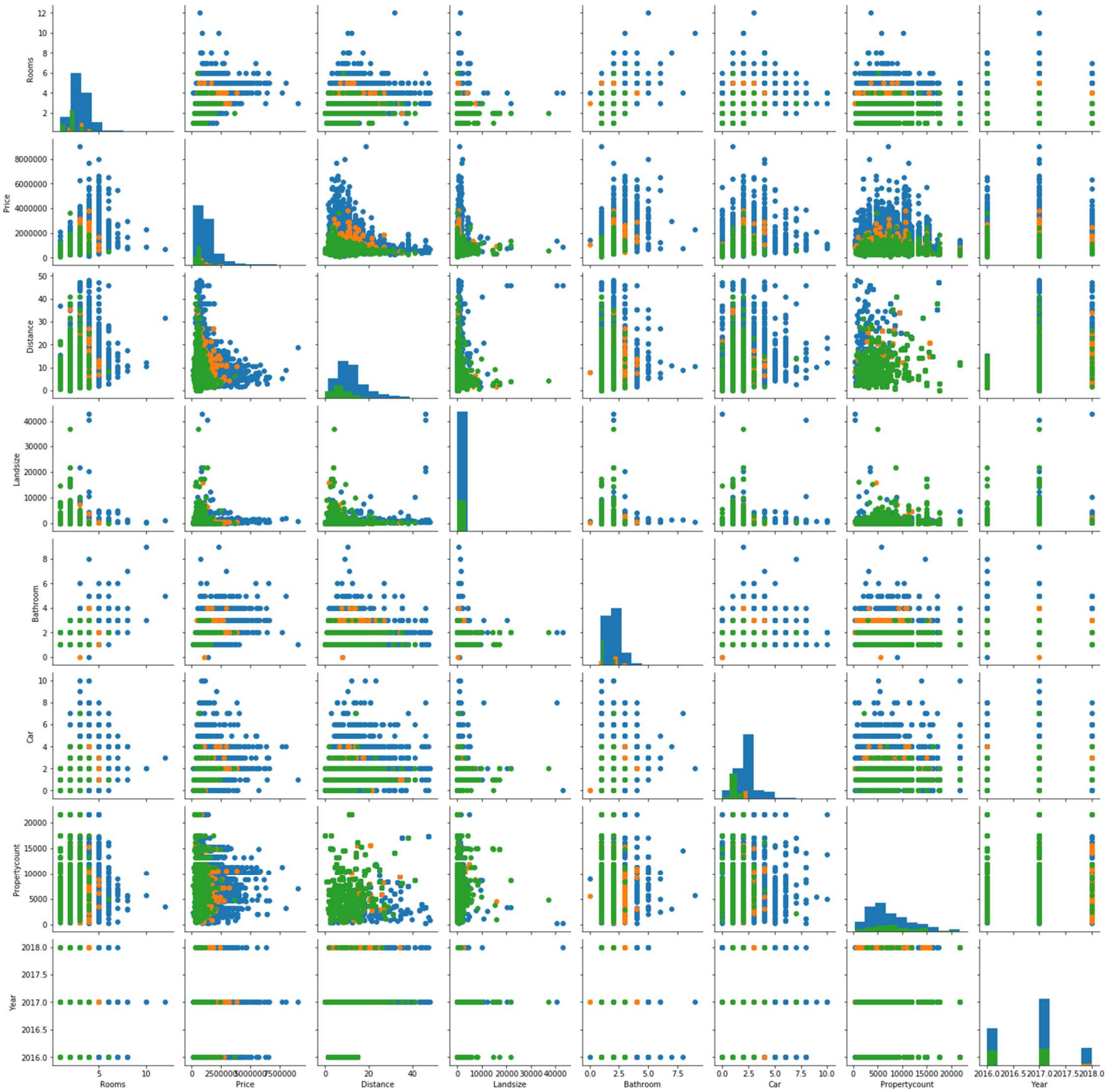
Histogram

This histogram is generated utilizing data frame df. This suggests mostly price is in range of $ 500,000 – 1,500,000.



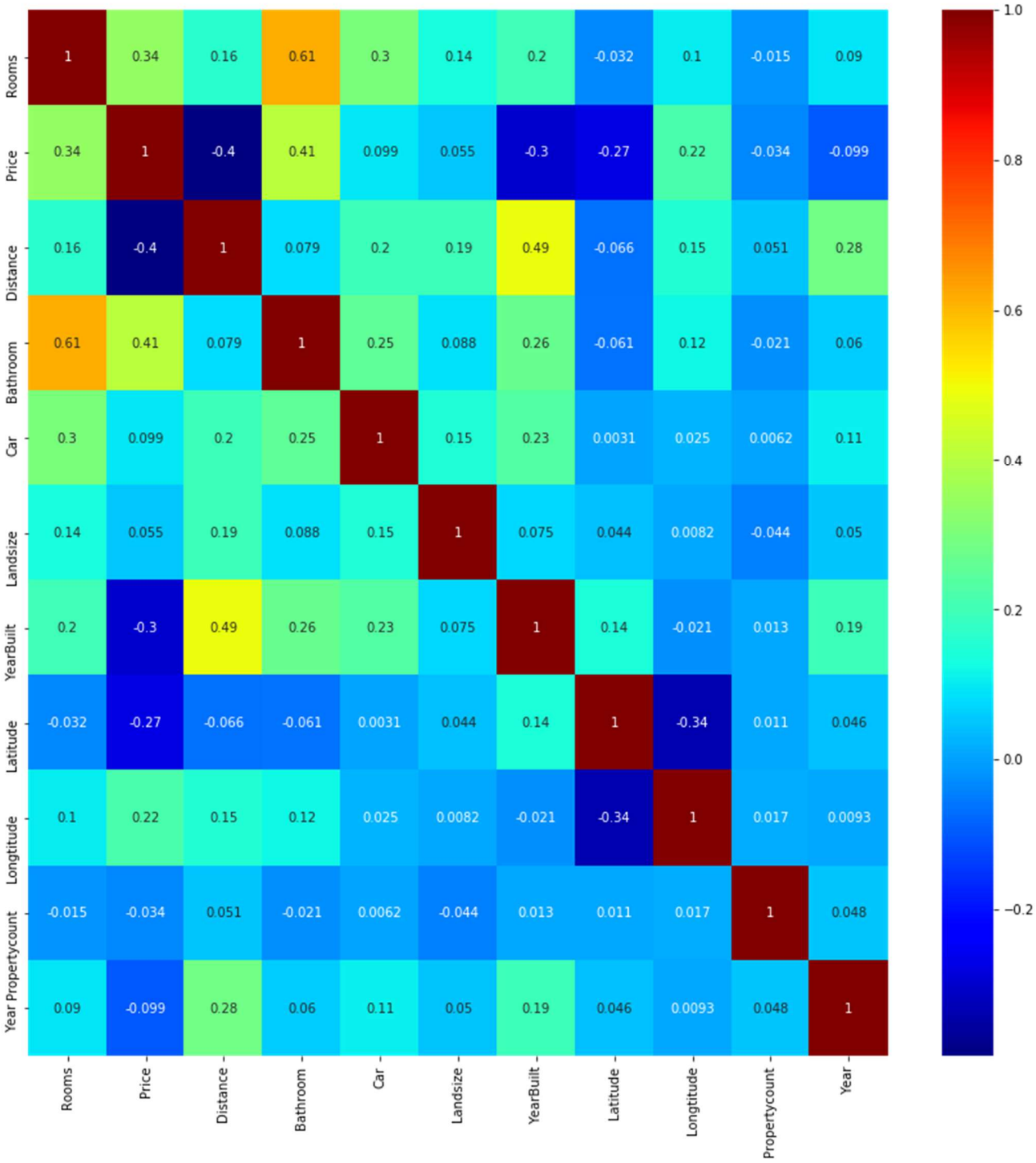
Pair plot/Pairgrid

Seaborn pairplot was prepared to see the inter-variable relationship between variables (Rooms, Price, Distance, Bathroom, Landsize, Car spot, Property count, Year & Type) but no direct linear relationship is observed as per plots.


Rooms
Price
Distance
Landsize
Bathroom
Car
Propertycount
Year

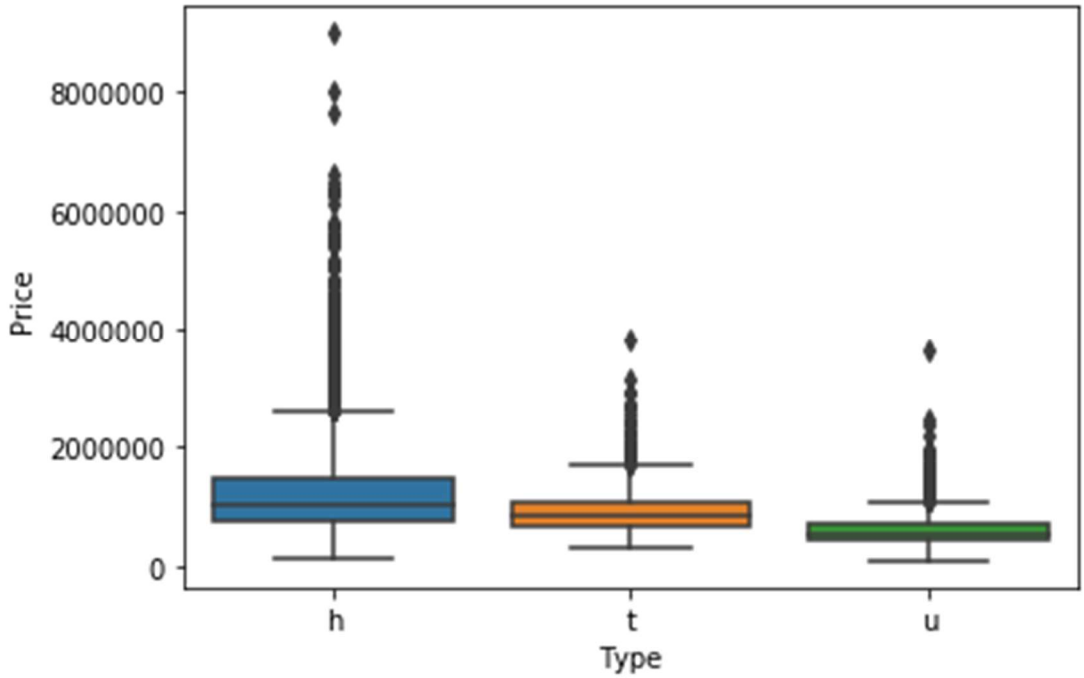
Heat Map

Correlation between numerical variables was visualized by heat map. Weak to strong correlation of price with distance, rooms, bathrooms & year built can be seen in heat map.



Box plot between price and type (house = h, unit = u & townhouse =t).

We see that the distributions of price between the different type of property have a significant overlap, and so Type would not be a good predictor of price.



Box plot between price and all regions.

Due to significant overlap between prices in all regions in Melbourne, regions can not be good predictor of price.

In order to visualize the box plot abbreviated form of regions was used as below;

Northern Metropolitan = N M

Western Metropolitan = W M

Southern Metropolitan = S M

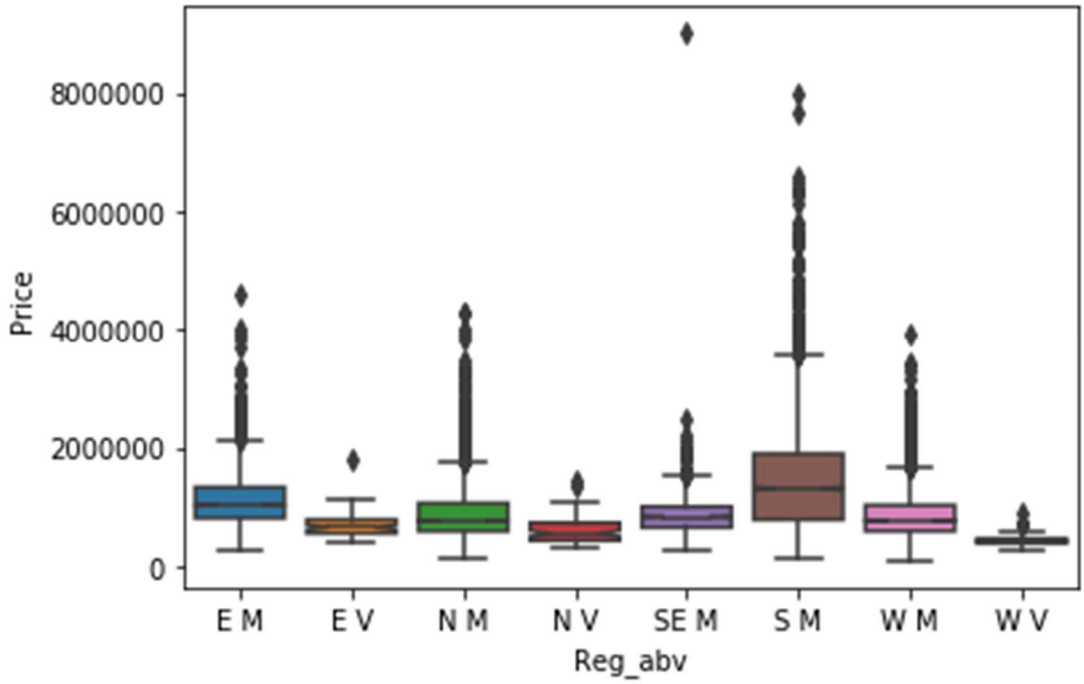
Eastern Metropolitan = E M

South-Eastern Metropolitan = SE M

Northern Victoria = N V

Eastern Victoria = E V

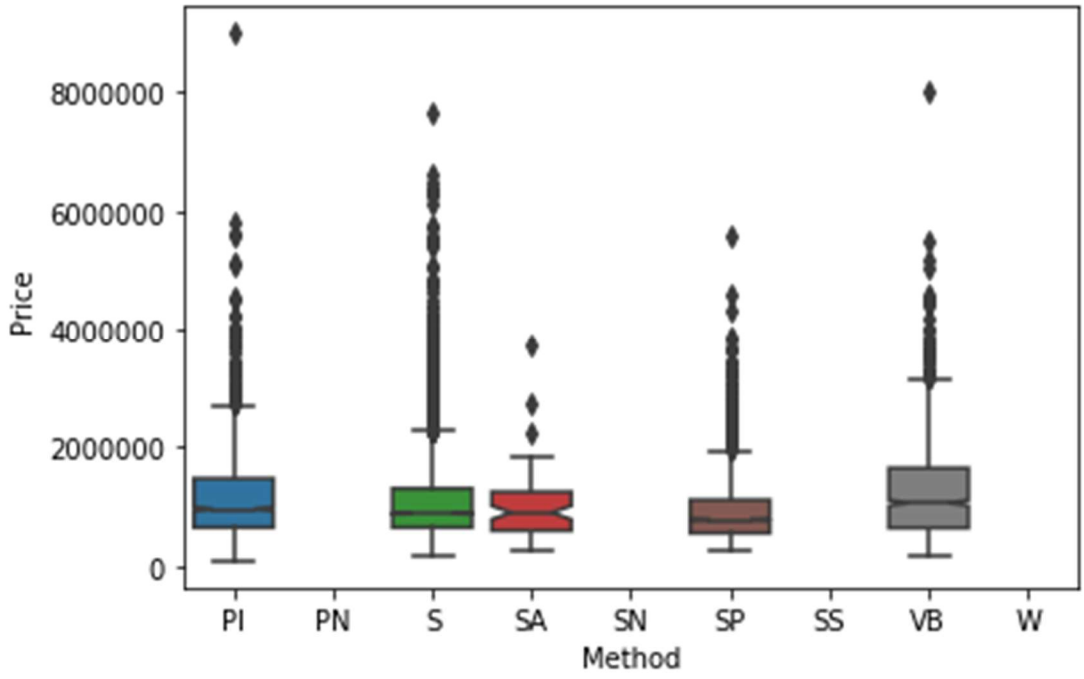
Western Victoria = W V



Box plot between price and method of selling.

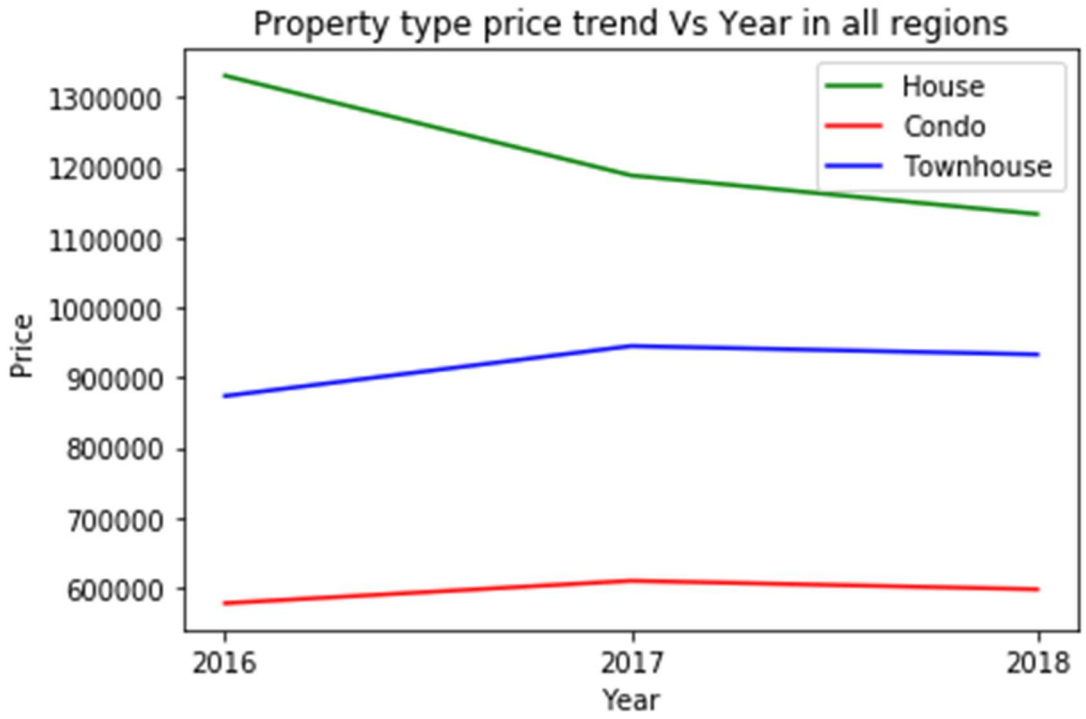
Plot suggest method is not a good indicator of price prediction.

S - property sold;
SP - property sold prior;
PI - property passed in;
PN - sold prior not disclosed;
SN - sold not disclosed;
VB - vendor bid;
W - withdrawn prior to auction;
SA - sold after auction;
SS - sold after auction price not disclosed.



Line plot of mean price of each type in Melbourne

Line plot between mean price of all three types of properties over the period of 3 years in Melbourne city. Graph suggests that House price decreased by ~ $100,000/year, Condo price climbed up slowly while Townhouse price remained almost same. It is time to built more condos in 2019 due to minimal change in price over 3 year period. Furthermore, it can be concluded that buyers should be interested in buying House in coming years due to dramatic change in price.



Model Development & Evaluation: Machine Learning Approach

Machine learning approach regression was chosen because of its simplicity and with the aid of Sklearn library implementation of model is quick and easy which is perfect to start the analyzing process. Regression approach was used to develop model for following dependent & independent variable;

- Number of bedrooms versus price
- Number of bathrooms versus price
- Year versus price
- Distance (from CBD) versus price
- Neighborhood venue count versus average property price
- Neighborhood venue category count versus average property price

The regression approach was used for above variables for;

1. All regions in Melbourne city
2. Southern Metropolitan (S M) Region
3. Northern Metropolitan (N M) Region

Regression approach is not limited to Linear type, it encompasses Multilinear, Polynomial, Ridge, Lasso Regression and finally Principal component analysis was sued for above 3 regions. Below is regression analysis performed for above mentioned relations with mentioned regression types.

All Regions in Melbourne City

Linear Regression

I. Number of rooms versus price for all property types in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.23 and mean squared error (MSE) of 3.474E+11. Price of property with 2 rooms was predicted, which found to be ~ $ 724427.

II. Year versus price for all property types in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0 and mean squared error (MSE) of 4.507E+11. Price of property in year 2019 was predicted, which found to be ~ $ 1081777.

III. Distance from Central build up area (CBD) versus price for all property types in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.05 and mean squared error (MSE) of 4.264E+11. Price of property 20 km from CBD was predicted, which found to be ~ $ 882389.

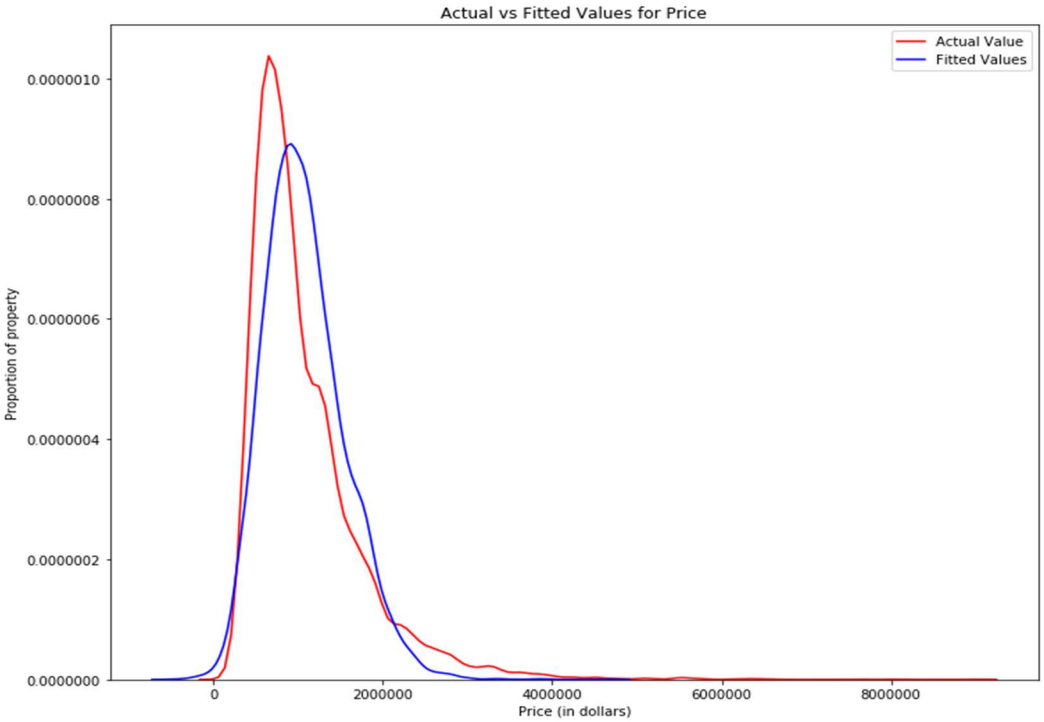
IV. Number of Bathroom versus price for all property types in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.22 and mean squared error (MSE) of 3.525E+11. Price of property with 2 bathrooms was predicted, which found to be ~ $ 1238825.

MultiLinear Regression (MLR) for all property types for all regions in Melbourne

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.5 and mean squared error (MSE) of 2.237E+11. Price of property with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 482900.

We can see that the fitted values are reasonably close to the actual values, since the two distributions overlap a bit. However, there is definitely some room for improvement.



Polynomial Regression with Pipe

In this case polynomial pipe regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.57 and mean squared error (MSE) of 1.949E+11. Price of property with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 444520.

Ridge Regression

In this case Ridge regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.55 and mean squared error (mse) of 1.859E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 0 to optimize R^2 value.

Lasso Regression

In this case Lasso regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.52 and mean squared error (mse) of 2.135E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 3 to optimize R^2 value.

Principal Component Analysis

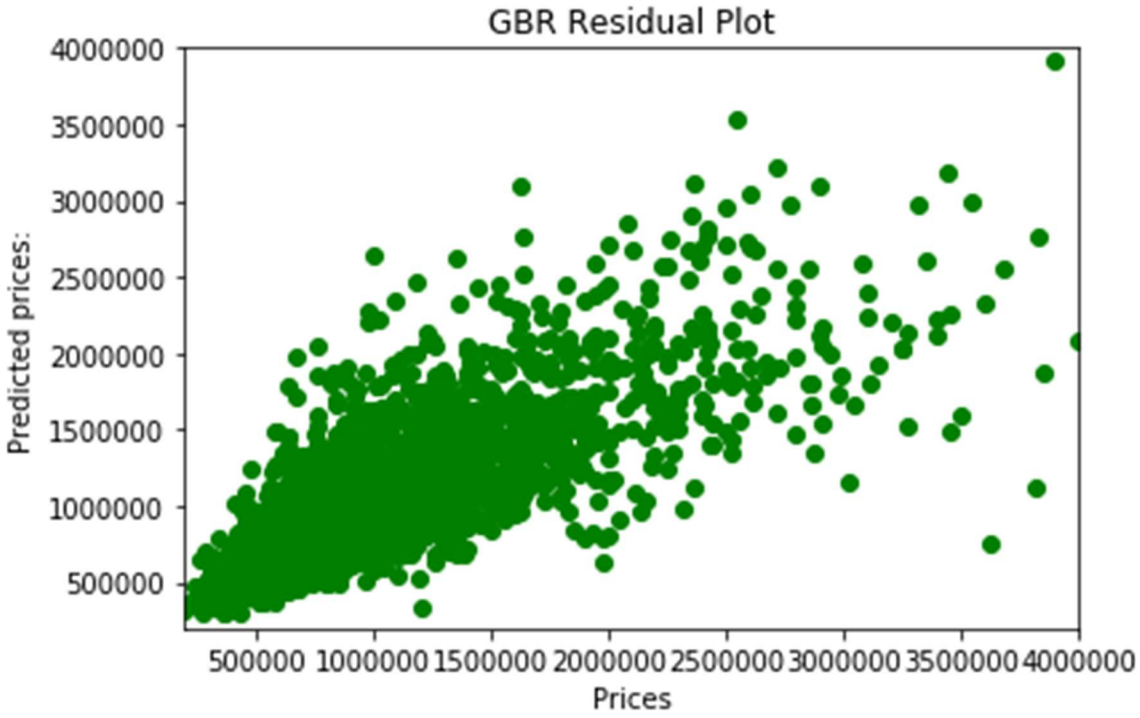
Principal component analysis (PCA), is a dimension reduction tool that projects data onto lower di- mensions, commonly referred to as principal components, in order to reduce the total number of variables to smaller data set with negligible information loss. In other words, if a feature is determined to be highly correlated to another, the feature is removed in order to help prevent overfitting of the model.

A heat map was constructed above confirming no notable correlation between the variables. In addition to this, a seaborn pairgrid plot was constructed to help visualise relationships between variables.

In this case Lasso regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.52 and mean squared error (mse) of 2.135E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 3 to optimize R^2 value.

Gradient Booster Regressor

In this case gradient booster regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.62 and mean squared error (MSE) of 1.587E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 0 and learning rate of 0.24 to optimize R^2 value.



All Region in Melbourne

Regression type	R^2	MAE	MSE	RMSE
Linear	0.23	407809	3.474E+11	589435
	0.05	456927	4.264E+11	653013
	0.22	425552	3.525E+11	593727
	0	479036	4.507E+11	671370
MLR	0.5	316917	2.237E+11	472982
Polynomial Pipe	0.57	290103	1.949E+11	441452
Ridge	0.55	284587	1.859E+11	431125
Lasso	0.52	308633	2.135E+11	462028
GBR	0.62	270466	1.587E+11	398392

While the lack of correlation was shown, PCA was tested in this model out of FOMO but it was confirmed to be detrimental to its overall performance.

Modeling for type = house in all Regions in Melbourne City

Linear Regression

I. Number of rooms versus price for type = house in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.12 and mean squared error (MSE) of 4.4435E+11. Price of house with 2 rooms was predicted, which found to be ~ $ 829940.

II. Year versus price for type = house in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.01 and mean squared error (MSE) of 4.9846E+11. Price of house in year 2019 was predicted, which found to be ~ $ 990582.

III. Distance from Central build up area (CBD) versus price for type = house in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.16 and mean squared error (MSE) of 4.2457E+11. Price of house 20 km from CBD was predicted, which found to be ~ $ 929938.

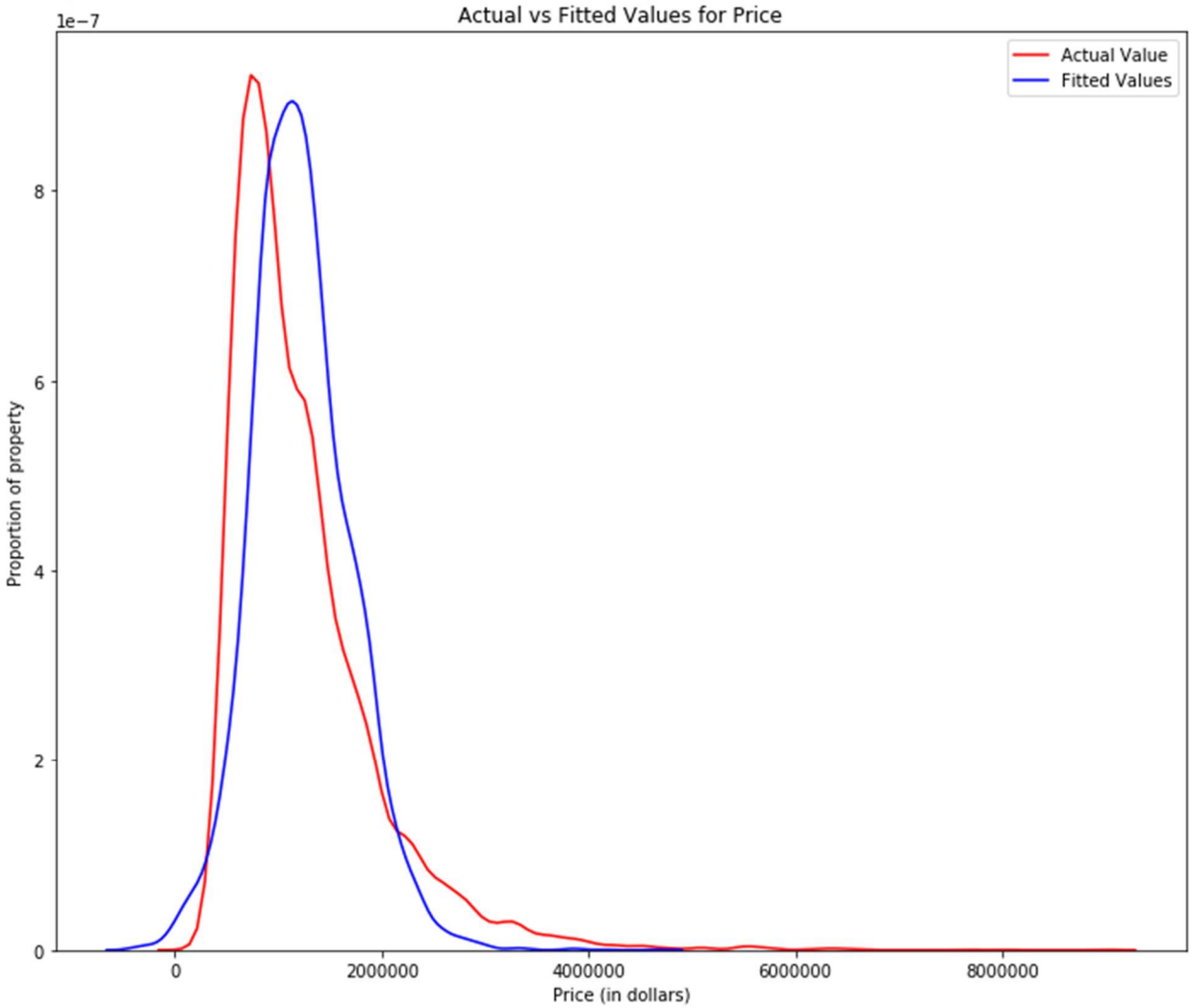
IV. Number of Bathroom versus price for type = house in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.17 and mean squared error (MSE) of 4.1831E+11. Price of house with 2 bathrooms was predicted, which found to be ~ $ 1330695.

MultiLinear Regression (MLR) for type = house for all regions in Melbourne

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.45 and mean squared error (MSE) of 2.7484E+11. Price of house with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 517153.

We can see that the fitted values are not reasonably close to the actual values, since the two distributions do not overlap a bit. However, there is definitely some room for improvement.



Polynomial Regression with Pipe for type = house in all regions in Melbourne

In this case polynomial pipe regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.53 and mean squared error (MSE) of 2.363E+11. Price of house with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 350900.

Ridge Regression for type = house in all regions in Melbourne

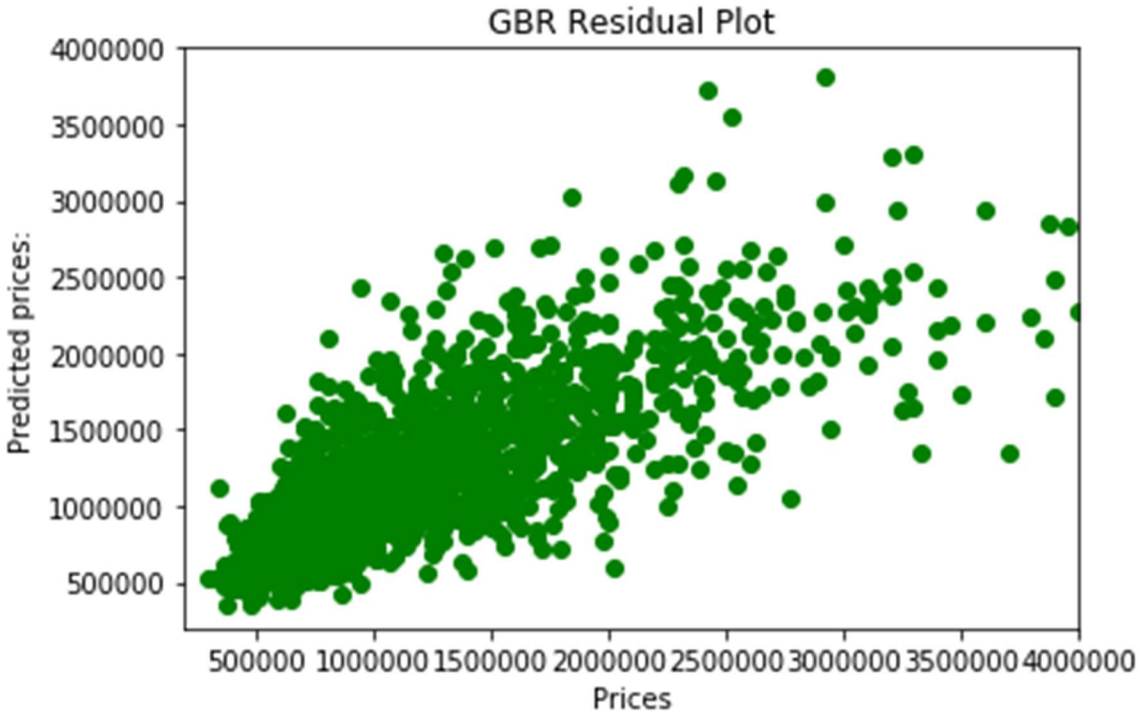
In this case Ridge regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.54 and mean squared error (MSE) of 2.4289E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 to optimize R^2 value.

Lasso Regression for type = house in all regions in Melbourne

In this case Lasso regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.46 and mean squared error (MSE) of 2.847E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 to optimize R^2 value.

Gradient Booster Regressor for type = house in all regions in Melbourne

In this case gradient booster regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.59 and mean squared error (MSE) of 2.19E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with learning rate of 0.19 to optimize R^2 value.



Summary of all regressions for type = house in all regions is tabulated in table below;

	Regression type	R^2	MAE	MSE	RMSE
House type in all regions Melbourne	Linear	0.12	482321	4.4435E+11	666598
		0.16	457295	4.2457E+11	651592
		0.17	468515	4.1831E+11	646768
		0.01	504354	4.9846E+11	706018
	MLR	0.45	359884	2.7484E+11	524248
	Polynomial Pipe	0.53	327849	2.3631E+11	486122
	Ridge	0.54	323986	2.4289E+11	492841
	Lasso	0.46	358092	2.8476E+11	533628
	GBR	0.59	308234	2.19E+11	467972

Modeling for type = unit in all Regions in Melbourne City

Linear Regression

I. Number of rooms versus price for type = unit in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.27 and mean squared error (MSE) of 4.43E+10. Price of unit with 2 rooms was predicted, which found to be ~ $ 601713.

II. Year versus price for type = unit in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0 and mean squared error (MSE) of 6E+10. Price of unit in year 2019 was predicted, which found to be ~ $ 669846.

III. Distance from Central build up area (CBD) versus price for type = unit in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0 and mean squared error (MSE) of 6E+10. Price of unit 20 km from CBD was predicted, which found to be ~ $ 558879.

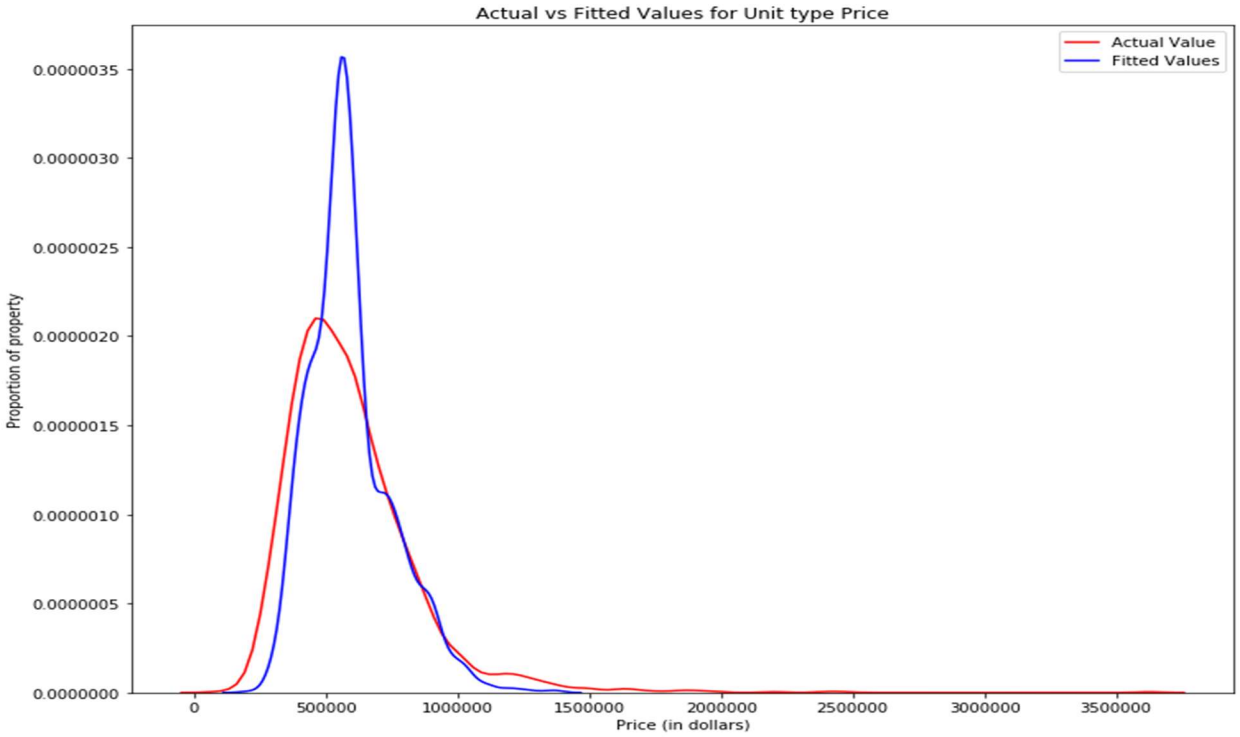
IV. Number of Bathroom versus price for type = unit in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.22 and mean squared error (MSE) of 4.68E+10. Price of unit with 2 bathrooms was predicted, which found to be ~ $ 830029.

MultiLinear Regression (MLR) for type = unit for all regions in Melbourne

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.42 and mean squared error (MSE) of 3.5E+10. Price of unit with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 666461.

We can see that the fitted values are not reasonably close to the actual values, since the two distributions do not overlap.



Polynomial Regression with Pipe for type = unit in all regions in Melbourne

In this case polynomial pipe regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.48 and mean squared error (MSE) of 3.13E+10. Price of unit with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 616668.

Ridge Regression for type = unit in all regions in Melbourne

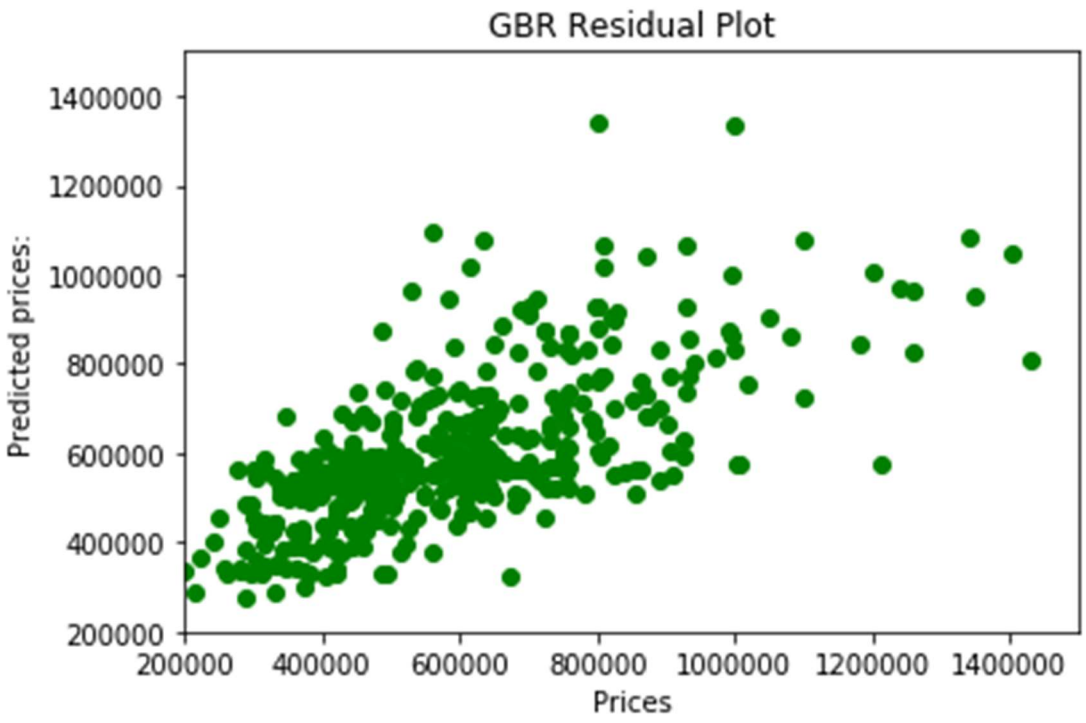
In this case Ridge regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.53 and mean squared error (MSE) of 2.71E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 to optimize R^2 value.

Lasso Regression for type = unit in all regions in Melbourne

In this case Lasso regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.39 and mean squared error (MSE) of 3.66E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 1 to optimize R^2 value.

Gradient Booster Regressor for type = unit in all regions in Melbourne

In this case gradient booster regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.49 and mean squared error (MSE) of 3.05E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with learning rate of 0.15 to optimize R^2 value.



Summary of all regressions for type = unit in all regions is tabulated in table below;

Unit type in all regions in Melbourne	Regression type	R^2	MAE	MSE	RMSE
	Linear	0.27	139612	4.43E+10	210393
		0	173580	6E+10	245034
		0.22	154869	4.68E+10	216363
		0	172374	6E+10	244951
	MLR	0.42	128069	3.5E+10	187039
	Polynomial Pipe	0.48	123458	3.13E+10	177005
	Ridge	0.53	123549	2.71E+10	164664
	Lasso	0.39	134973	3.66E+10	191342
	GBR	0.49	126081	3.05E+10	174704

Modeling for type = Townhouse in all Regions in Melbourne City

Linear Regression

I. Number of rooms versus price for type = townhouse in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.19 and mean squared error (MSE) of 1.219E+11. Price of townhouse with 2 rooms was predicted, which found to be ~ $ 701027.

II. Year versus price for type = townhouse in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.01 and mean squared error (MSE) of 1.495E+11. Price of townhouse in year 2019 was predicted, which found to be ~ $ 1029662.

III. Distance from Central build up area (CBD) versus price for type = townhouse in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of .02 and mean squared error (MSE) of 1.474E+11. Price of townhouse 20 km from CBD was predicted, which found to be ~ $ 808019.

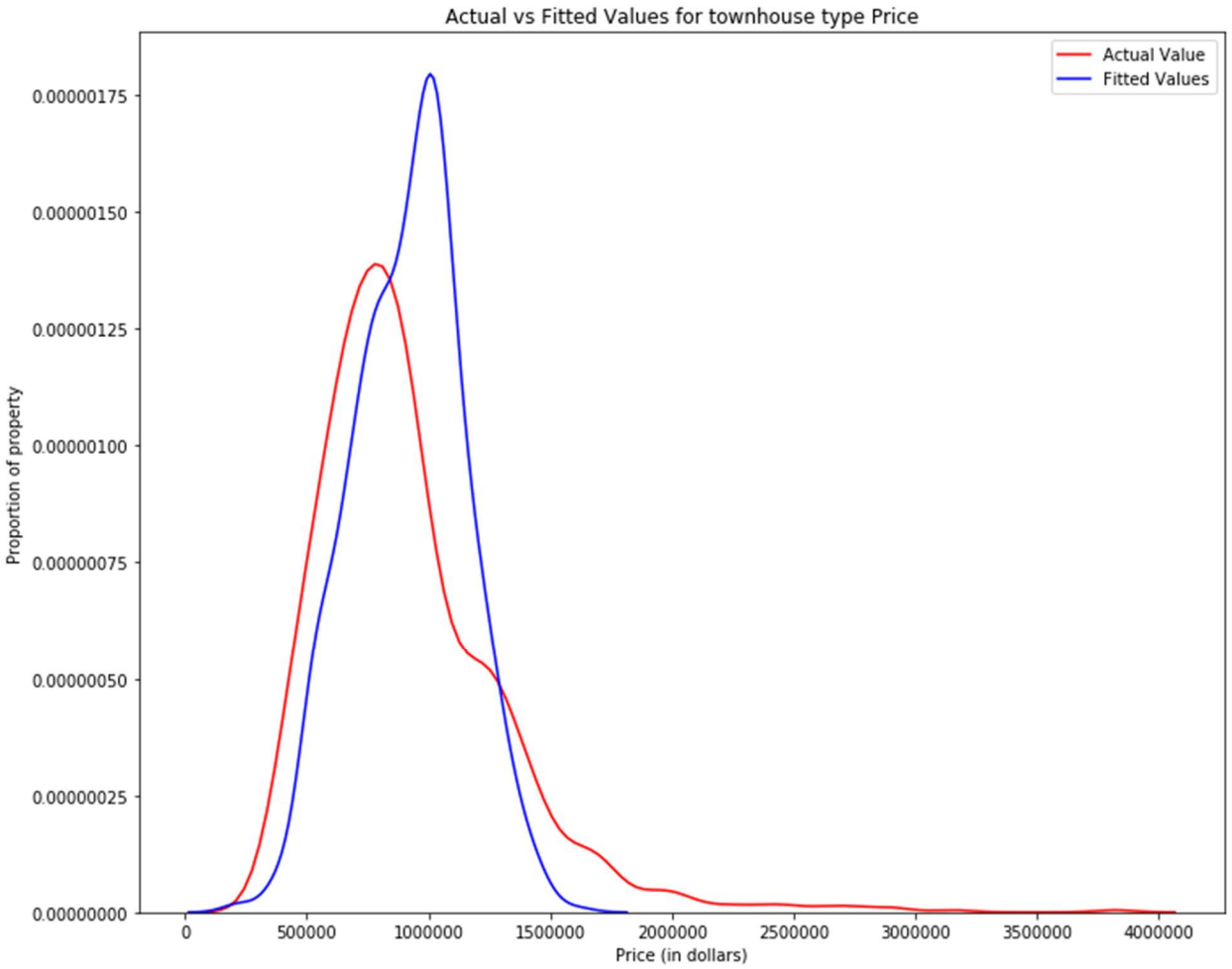
IV. Number of Bathroom versus price for type = townhouse in all regions

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.2 and mean squared error (MSE) of 1.2E+11. Price of townhouse with 2 bathrooms was predicted, which found to be ~ $ 968423.

MultiLinear Regression (MLR) for type = townhouse for all regions in Melbourne

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.34 and mean squared error (MSE) of 9.85E+10. Price of townhouse with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 641096.

We can see that the fitted values are not reasonably close to the actual values, since the two distributions do not overlap.



Polynomial Regression with Pipe for type = townhouse in all regions in Melbourne
In this case polynomial pipe regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.4 and mean squared error (MSE) of 8.96E+10. Price of townhouse with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 558819.

Ridge Regression for type = townhouse in all regions in Melbourne

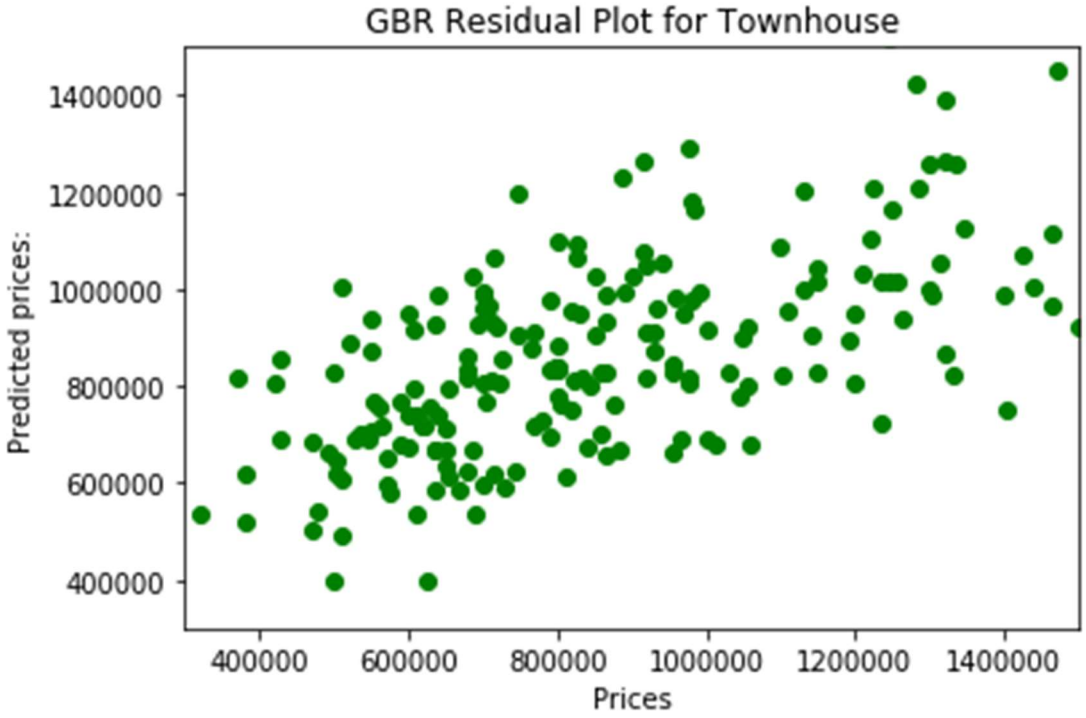
In this case Ridge regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.39 and mean squared error (MSE) of 9.58E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 to optimize R^2 value.

Lasso Regression for type = townhouse in all regions in Melbourne

In this case Lasso regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.35 and mean squared error (MSE) of 1.01E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 to optimize R^2 value.

Gradient Booster Regressor for type = townhouse in all regions in Melbourne

In this case gradient booster regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of 0.45 and mean squared error (MSE) of 8.8E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with learning rate of 0.15 to optimize R^2 value.



Summary of all regressions for type = townhouse in all regions is tabulated in table below;

Townhouse type in all regions in Melbourne	Regression type	R^2	MAE	MSE	RMSE
	Linear	0.19	244730	1.2193E+11	349186
		0.02	279363	1.4749E+11	384038
		0.2	245754	1.2008E+11	346522
		0.01	279083	1.4955E+11	386711
	MLR	0.34	221245	9.8567E+10	313953
	Polynomial Pipe	0.4	209798	8.9632E+10	299386
	Ridge	0.39	223910	9.5873E+10	309634
	Lasso	0.35	234859	1.0189E+11	319204
	GBR	0.45	210631	8.8098E+10	296813

Southern Metropolitan (S M) Region

Linear Regression

I. Number of rooms versus price for all property types in S M Region

In this case Linear regression was performed utilizing data frame df_S_M, which resulted in regression coefficient (R^2) of 0.49 and mean squared error (MSE) of 4.058E+11. Price of property with 2 rooms was predicted, which found to be ~ $ 886874.

II. Year versus price for all property types in S M Region

In this case Linear regression was performed utilizing data frame df_S_M, which resulted in regression coefficient (R^2) of 0.02 and mean squared error (MSE) of 7.787E+11. Price of property in year 2019 was predicted, which found to be ~ $ 1961055.

III. Distance from Central build up area (CBD) versus price for all property types in S M Region

In this case Linear regression was performed utilizing data frame df_S_M, which resulted in regression coefficient (R^2) of 0 and mean squared error (MSE) of 7.945E+11. Price of property 20 km from CBD was predicted, which found to be ~ $ 1365098.

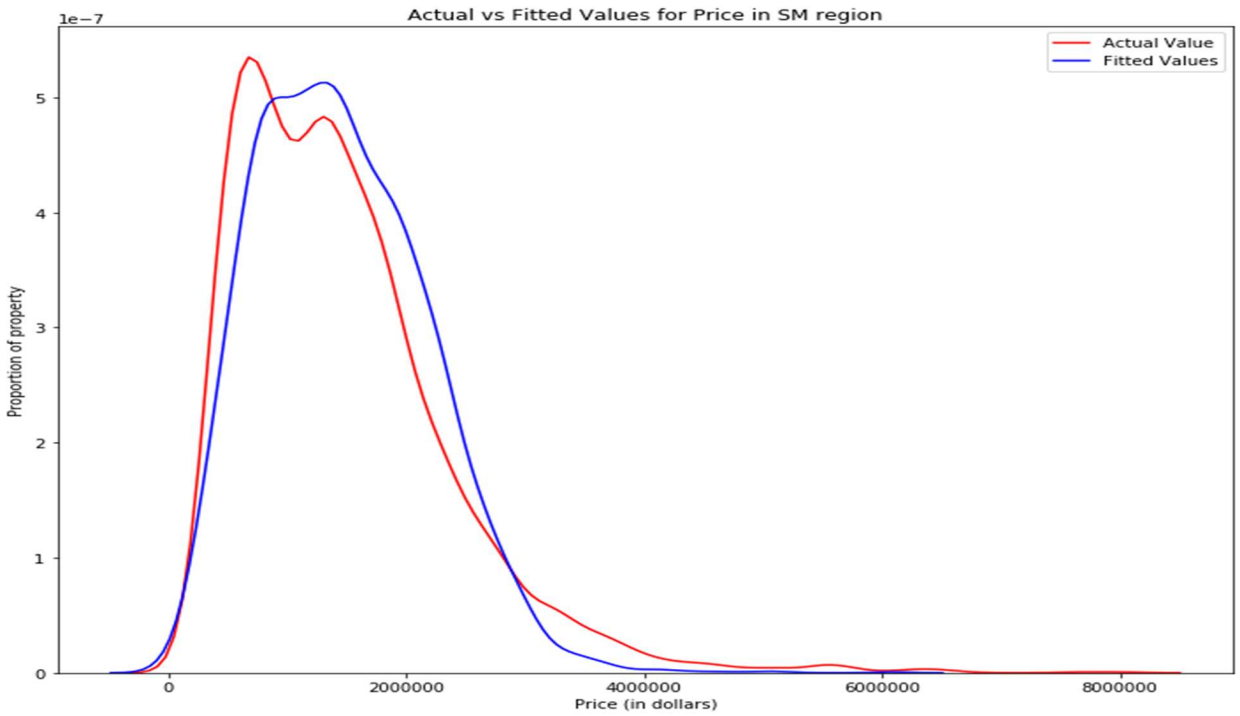
IV. Number of Bathroom versus price for all property types in S M Region

In this case Linear regression was performed utilizing data frame df_S_M, which resulted in regression coefficient (R^2) of 0.39 and mean squared error (MSE) of 4.87E+11. Price of property with 2 bathrooms was predicted, which found to be ~ $ 1651077.

MultiLinear Regression (MLR) for all property types in S M Region

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df_SM, which resulted in regression coefficient (R^2) of 0.63 and mean squared error (MSE) of 2.92E+11. Price of property with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 522907.

We can see that the fitted values are reasonably close to the actual values, since the two distributions overlap a bit. However, there is definitely some room for improvement.



Polynomial Regression with Pipe for all property types in S M Region

In this case polynomial pipe regression was performed utilizing data frame df_S_M, which resulted in regression coefficient (R^2) of 0.68 and mean squared error (MSE) of 2.509E+11. Price of property with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 627889.

Ridge Regression for all property types in S M Region

In this case Ridge regression was performed utilizing data frame df_S_M, which resulted in regression coefficient (R^2) of 0.69 and mean squared error (MSE) of 2.807E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 to optimize R^2 value.

Lasso Regression for all property types in S M Region

In this case Lasso regression was performed utilizing data frame df_S_M, which resulted in regression coefficient (R^2) of 0.63 and mean squared error (MSE) of 2.893E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 0 to optimize R^2 value.

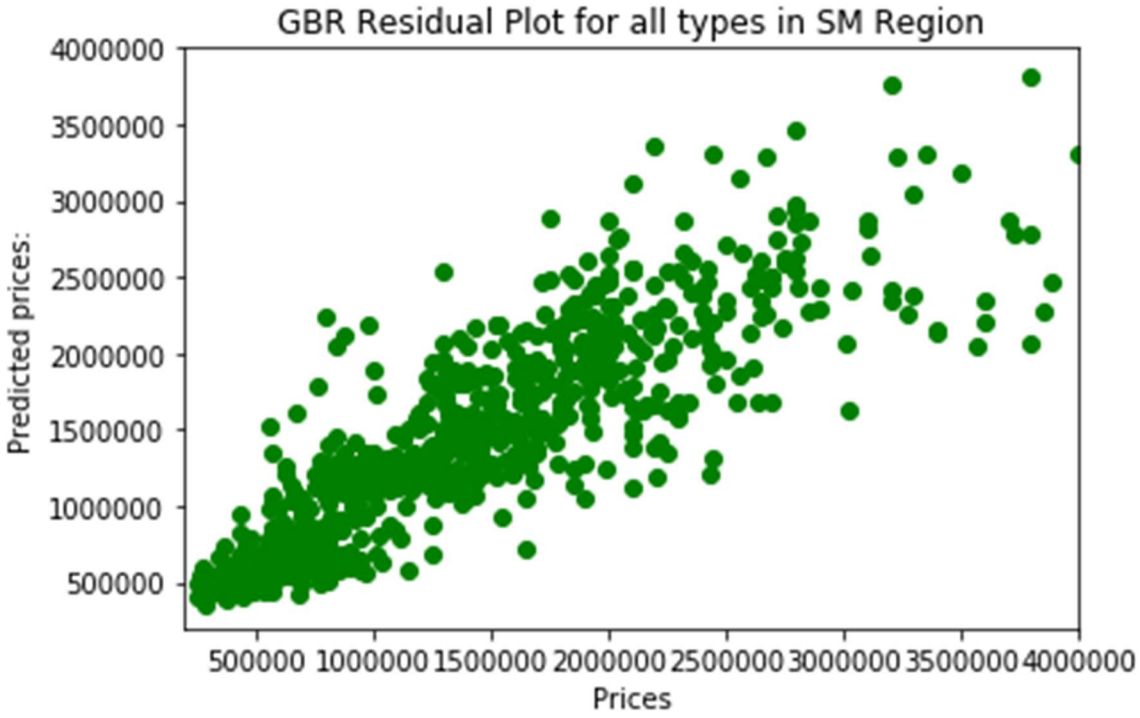
Principal Component Analysis

Principal component analysis (PCA), is a dimension reduction tool that projects data onto lower di- mensions, commonly referred to as principal components, in order to reduce the total number of variables to smaller data set with negligible information loss. In other words, if a feature is determined to be highly correlated to another, the feature is removed in order to help prevent overfitting of the model.

A heat map was constructed above confirming no notable correlation between the variables. In addition to this, a seaborn pairgrid plot was constructed to help visualise relationships between variables.

Gradient Booster Regressor

In this case gradient booster regression was performed utilizing data frame df_S_M, which resulted in regression coefficient (R^2) of 0.75 and mean squared error (MSE) of 2.317E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 and learning rate of 0.15 to optimize R^2 value.



Summary of all regressions for all type of property in Southern Metropolitan Region is tabulated in table below;

	Regression type	R^2	MAE	MSE	RMSE
Southern Metropolitan Region	Linear	0.49	438260	4.058E+11	637025
		0	669042	7.945E+11	891349
		0.39	517633	4.871E+11	697946
		0.02	662206	7.787E+11	882458
	MLR	0.63	365962	2.924E+11	540745
	Polynomial Pipe	0.68	338543	2.509E+11	500889
	Ridge	0.69	353303	2.807E+11	529781
	Lasso	0.63	362870	2.893E+11	537895
	GBR	0.75	314712	2.317E+11	481312

Modeling for type = house in Southern Metropolitan (S M) Region

Linear Regression

I. Number of rooms versus price for type = house in S M Region

In this case Linear regression was performed utilizing data frame df_S_M_h, which resulted in regression coefficient (R^2) of 0.26 and mean squared error (MSE) of 5.376E+11. Price of house with 2 rooms was predicted, which found to be ~ $ 1197267.

II. Year versus price for type = house in S M Region

In this case Linear regression was performed utilizing data frame df_S_M_h, which resulted in regression coefficient (R^2) of 0.01 and mean squared error (MSE) of 7.19E+11. Price of house in year 2019 was predicted, which found to be ~ $ 2170160.

III. Distance from Central build up area (CBD) versus price for type = house in S M Region

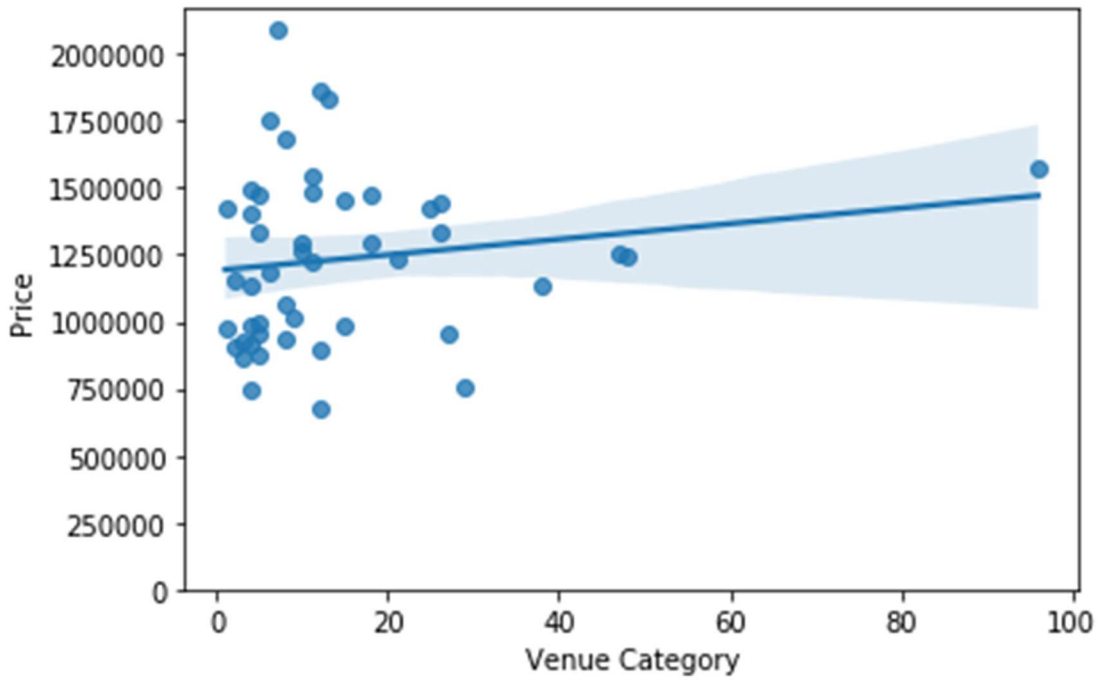
In this case Linear regression was performed utilizing data frame df_S_M_h, which resulted in regression coefficient (R^2) of 0.06 and mean squared error (MSE) of 6.79E+11. Price of house 20 km from CBD was predicted, which found to be ~ $ 1274496.

IV. Number of Bathroom versus price for type = house in S M Region

In this case Linear regression was performed utilizing data frame df_S_M_h, which resulted in regression coefficient (R^2) of 0.3 and mean squared error (MSE) of 5.1E+11. Price of house with 2 bathrooms was predicted, which found to be ~ $ 1930511.

V. Venue category count versus price for type = house in S M Region

In this case seaborn plot was generated to investigate if linear relationship exists between venue count and average property price. Seaborn suggested poor relationship between venue count & average property price which was supplemented by linear regression model which computed regression coefficient(R^2) of 0.0 & mean squared error of 3.4E+11



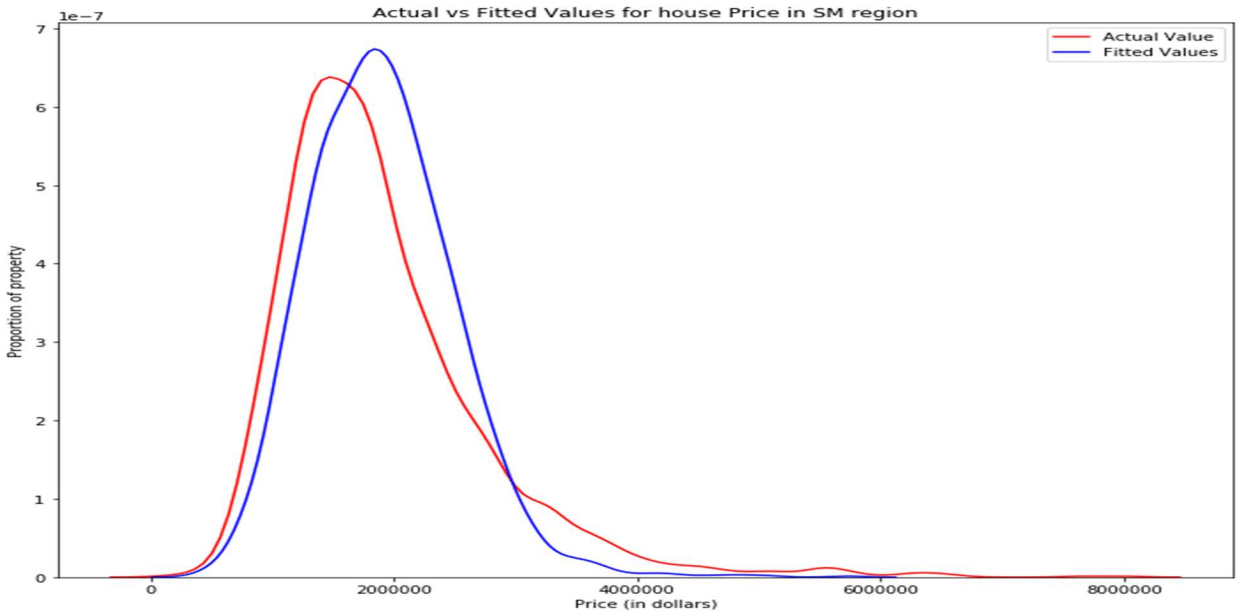
VI. Neighborhood venue category versus average property price for type = house in S M Region

In order to establish relationship between neighborhood venue category versus average property price, ve nue category from data frame south_metropolitan_h_venues were converted into binary format 0 & 1 using one hot encoding which was defined as a new data frame by the name of south_metropolitan_h_onehot. Nei ghborhood from south_metropolitan_h_venues was inserted into data frame north_metropolitan_h_onehot. Ve nue category in south_metropolitan_h_onehot was grouped & summed based on Neighbourhood and was de fined to a new data frame joined_venues_price_S_M_h_df. which was utilized for regression modelling whi ch resulted in regression coefficient (R^2) of ~ -2.3 and mean squared error (mse) of 8.11E+11.

MultiLinear Regression (MLR) for type = house in S M Region

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df_S_M_h, which resulted in regression coefficient (R^2) of 0.48 and mean squared error (MSE) of 3.75E+11. Price of house with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 532558.

We can see that the fitted values are reasonably close to the actual values, since the two distributions overlap a bit. However, there is definitely some room for improvement.



Polynomial Regression with Pipe for type = house in S M Region

In this case polynomial pipe regression was performed utilizing data frame df_S_M_h, which resulted in regression coefficient (R^2) of 0.61 and mean squared error (MSE) of 2.85E+11. Price of house with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 870663.

Ridge Regression for type = house in S M Region

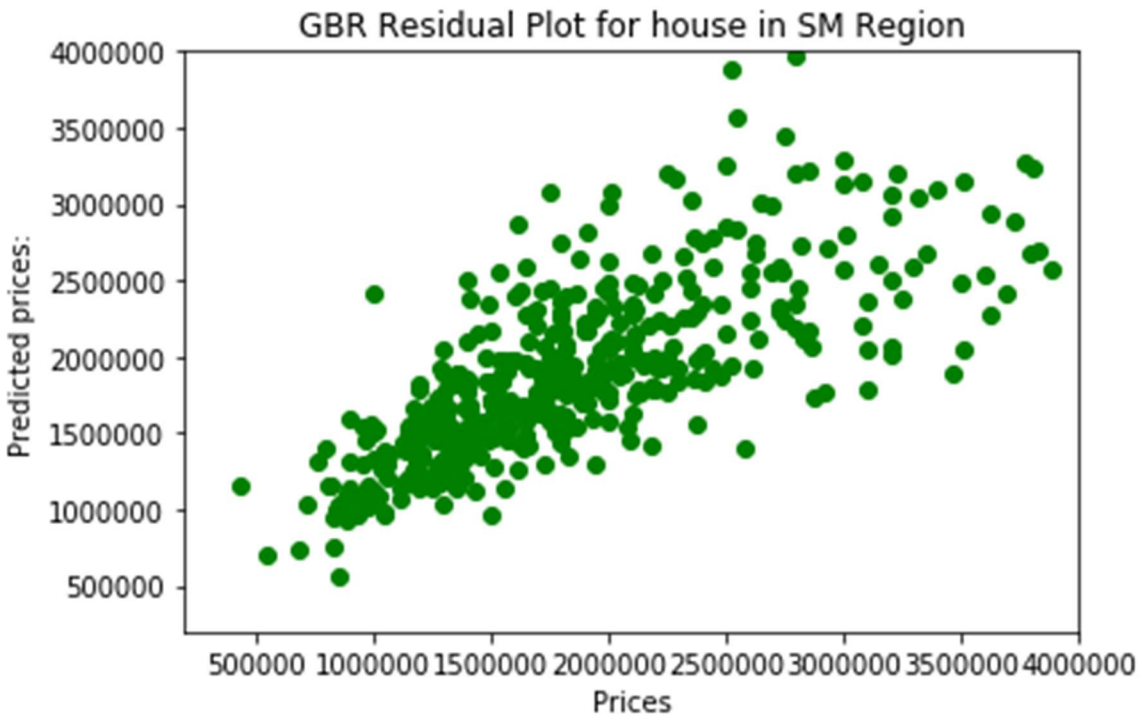
In this case Ridge regression was performed utilizing data frame df_S_M_h, which resulted in regression coefficient (R^2) of 0.61 and mean squared error (MSE) of 3.01E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 0 to optimize R^2 value.

Lasso Regression for type = house in S M Region

In this case Lasso regression was performed utilizing data frame df_S_M_h, which resulted in regression coefficient (R^2) of 0.48 and mean squared error (MSE) of 2.68E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 3 to optimize R^2 value.

Gradient Booster Regressor for type = house in S M Region

In this case gradient booster regression was performed utilizing data frame df_S_M_h, which resulted in regression coefficient (R^2) of 0.61 and mean squared error (MSE) of 3.04E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 1 and learning rate of 0.1 to optimize R^2 value.

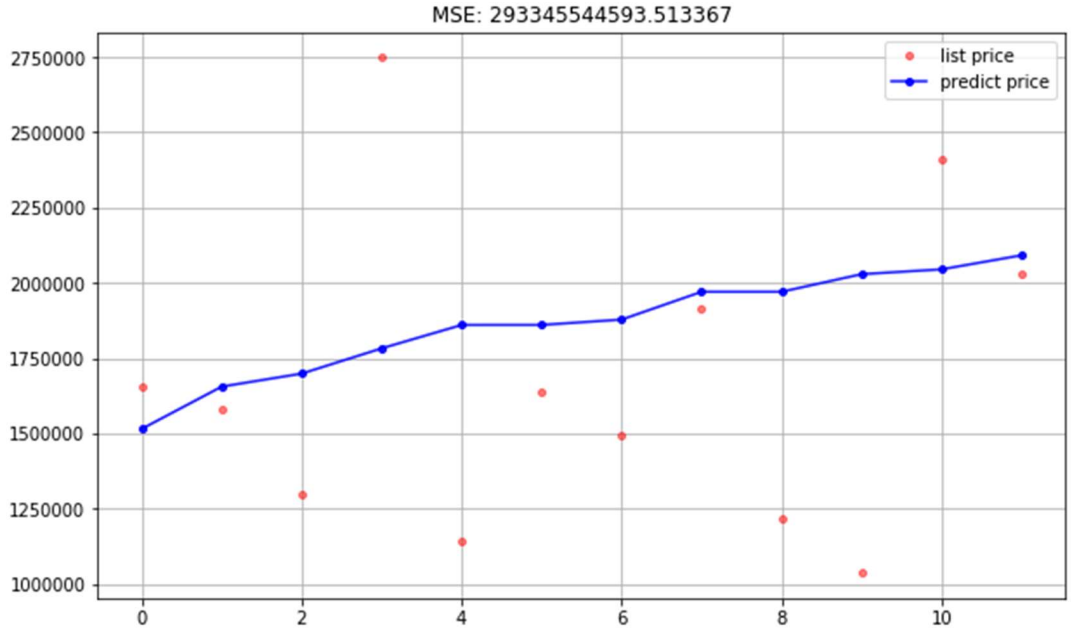


Summary of all regressions for type = house in Southern Metropolitan Region is tabulated in table below;

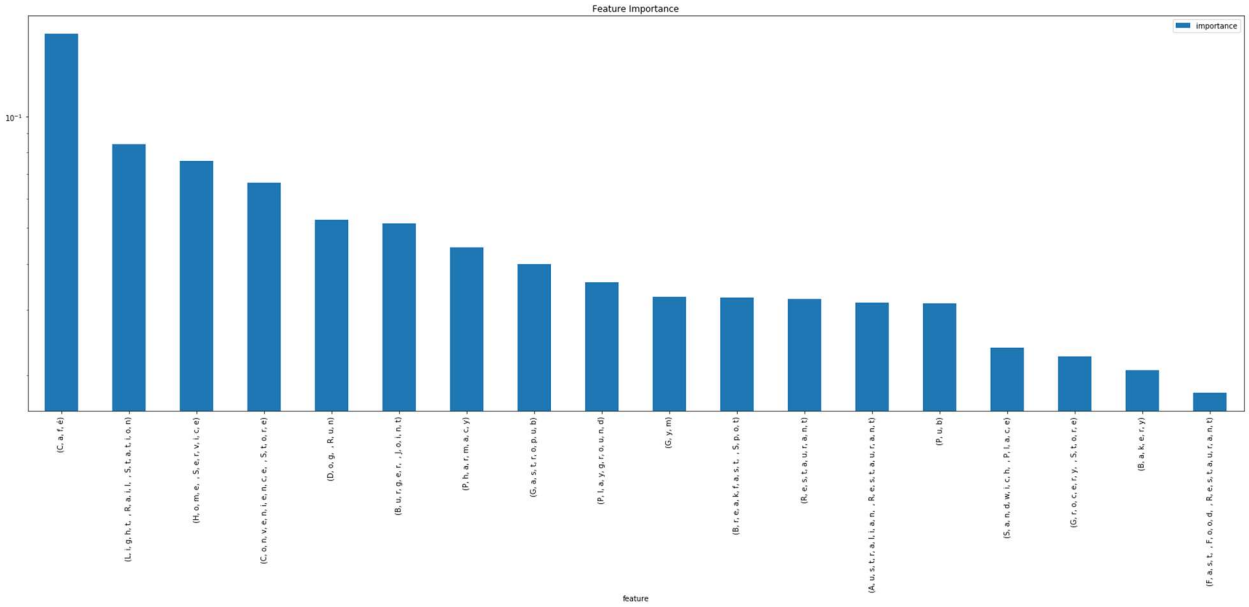
	Regression type	R^2	MAE	MSE	RMSE
House Type Southern Metropolitan Region	Linear	0.26	513922	5.376E+11	733212
		0.06	589097	6.7955E+11	824346
		0.3	505210	5.1025E+11	714320
		0.01	610849	7.1993E+11	848487
	MLR	0.48	422645	3.7598E+11	613172
	Polynomial Pipe	0.61	371753	2.858E+11	534598
	Ridge	0.61	388576	3.0172E+11	549292
	Lasso	0.48	386729	2.6819E+11	517867
	GBR	0.61	376565	3.0421E+11	551548

Random Forest Regression for type = house in S M Region

Random Forest Regression was carried out to determine relationship between venue category and average property price. In this regard data frame joined_venues_price_S_M_h_df was utilized which resulted in regression coefficient (R^2) of ~ -0.19 with mean squared error (mse) of ~ 2.93E+11. Grid Search CV was utilized to determine best parameters including max_depth, min_samples_leaf & n_estimators in order to optimize R^2 value. Below plot shows list price versus predicted price;



In addition to this, feature importance using random forest regression was performed to determine venues with respect to their importance on average property prices of neighbourhoods. Below is the plot showing venue with its importance with respect to average property price;



Modeling for type = unit in Southern Metropolitan (S M) Region
Linear Regression

I. Number of rooms versus price for type = unit in S M Region

In this case Linear regression was performed utilizing data frame df_S_M_u, which resulted in regression coefficient (R^2) of 0.36 and mean squared error (MSE) of 4.74E+10. Price of unit with 2 rooms was predicted, which found to be ~ $ 676644.

II. Year versus price for type = unit in S M Region

In this case Linear regression was performed utilizing data frame df_S_M_u, which resulted in regression coefficient (R^2) of 0.01 and mean squared error (MSE) of 7.41E+10. Price of unit in year 2019 was predicted, which found to be ~ $ 756801.

III. Distance from Central build up area (CBD) versus price for type = unit in S M Region

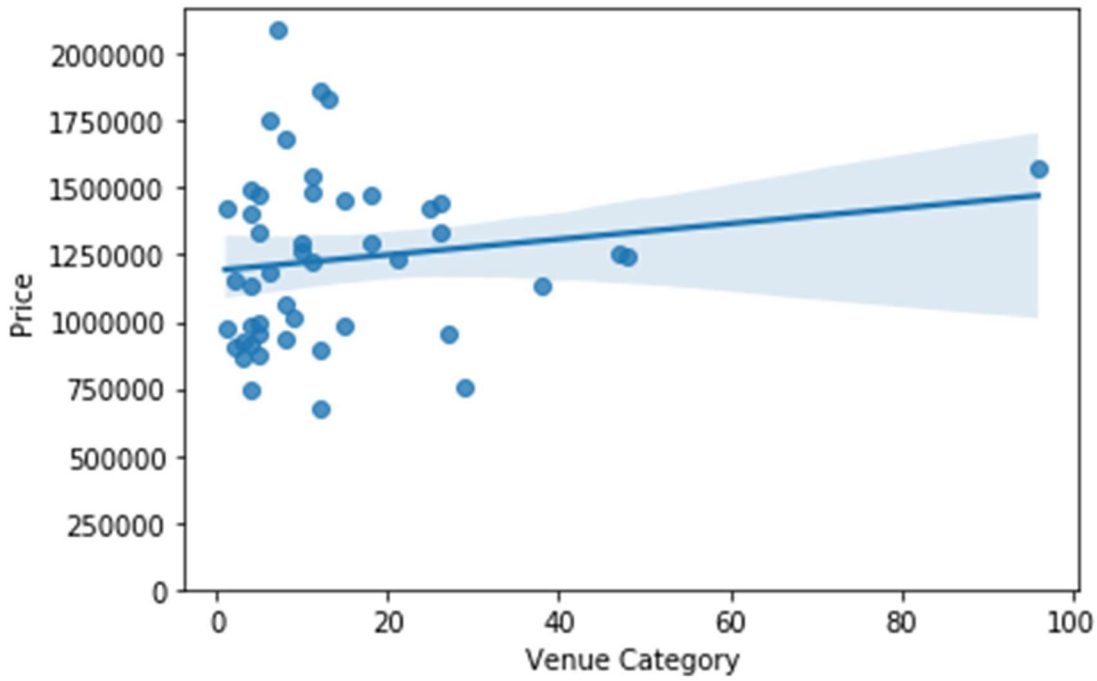
In this case Linear regression was performed utilizing data frame df_S_M_u, which resulted in regression coefficient (R^2) of 0 and mean squared error (MSE) of 7.45E+10. Price of unit 20 km from CBD was predicted, which found to be ~ $ 661398.

IV. Number of Bathroom versus price for type = unit in S M Region

In this case Linear regression was performed utilizing data frame df_S_M_u, which resulted in regression coefficient (R^2) of 0.25 and mean squared error (MSE) of 5.6E+10. Price of unit with 2 bathrooms was predicted, which found to be ~ $ 924835.

V. Venue category count versus price for type = unit in S M Region

In this case seaborn plot was generated to investigate if linear relationship exists between venue count and average property price. Seaborn suggested poor relationship between venue count & average property price which was supplemented by linear regression model which computed regression coefficient(R^2) of 0.0 & mean squared error of 1.27E+10.



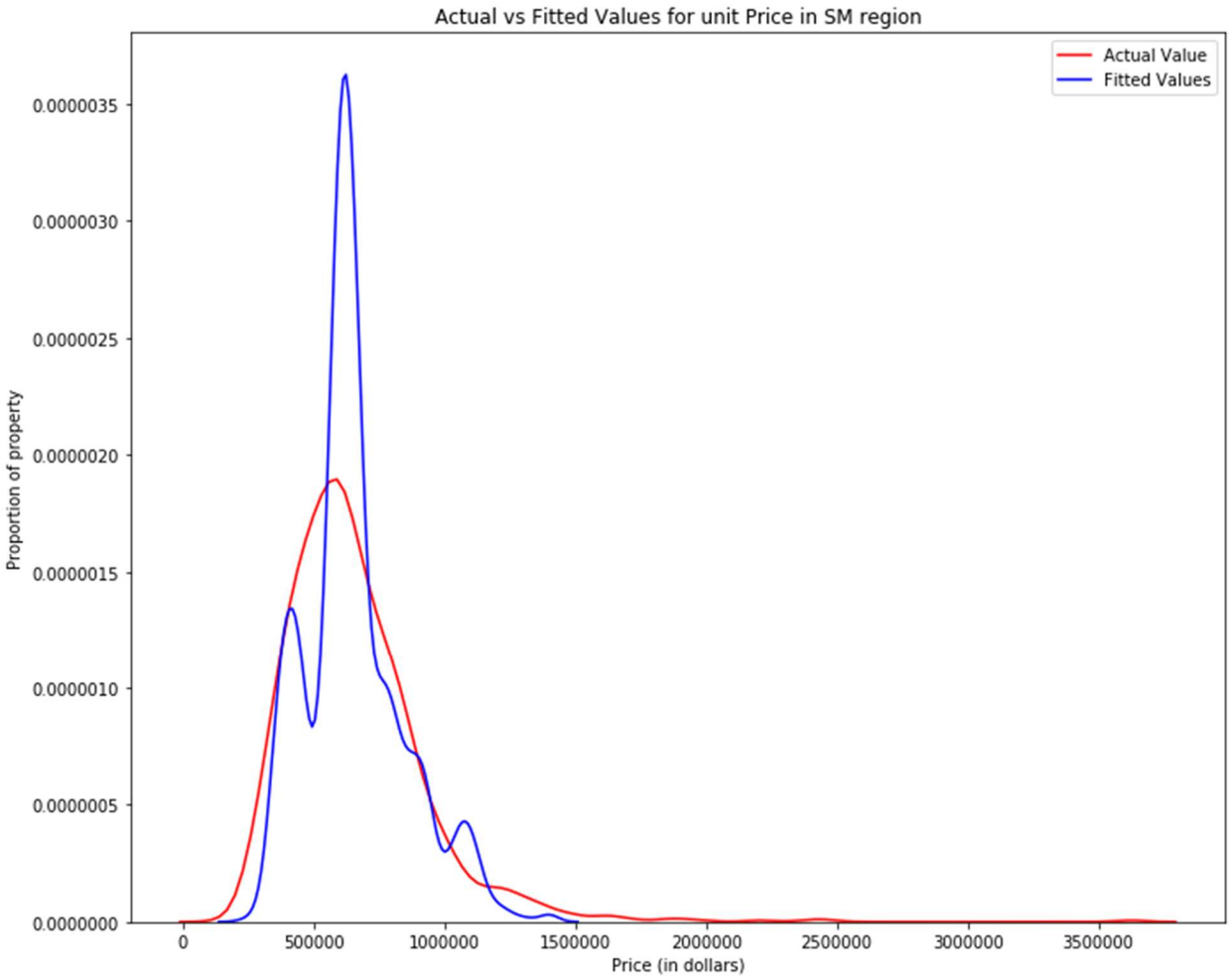
VI. Neighborhood venue category versus average property price for type = unit in S M Region

In order to establish relationship between neighborhood venue category versus average property price, ve nue category from data frame south_metropolitan_u_venues were converted into binary format 0 & 1 using one hot encoding which was defined as a new data frame by the name of south_metropolitan_t_onehot. Neig hborhood from south_metropolitan_u_venues was inserted into data frame north_metropolitan_t_onehot. Ven ue category in south_metropolitan_u_onehot was grouped & summed based on Neighbourhood and was defi ned to a new data frame joined_venues_price_S_M_u_df. which was utilized for regression modelling which resulted in regression coefficient (R^2) of ~ -3.5 and mean squared error (mse) of 3.2E+10.

MultiLinear Regression (MLR) for type = unit in S M Region

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df_S_M_u, which resulted in regression coefficient (R^2) of 0.49 and mean squared error (MSE) of 3.8E+10. Price of unit with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 770772.

We can see that the fitted values are not reasonably close to the actual values, since the two distributions do not overlap.



Polynomial Regression with Pipe for type = unit in S M Region

In this case polynomial pipe regression was performed utilizing data frame df_S_M_u, which resulted in regression coefficient (R^2) of 0.53 and mean squared error (MSE) of 3.48E+10. Price of unit with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 925145.

Ridge Regression for type = unit in S M Region

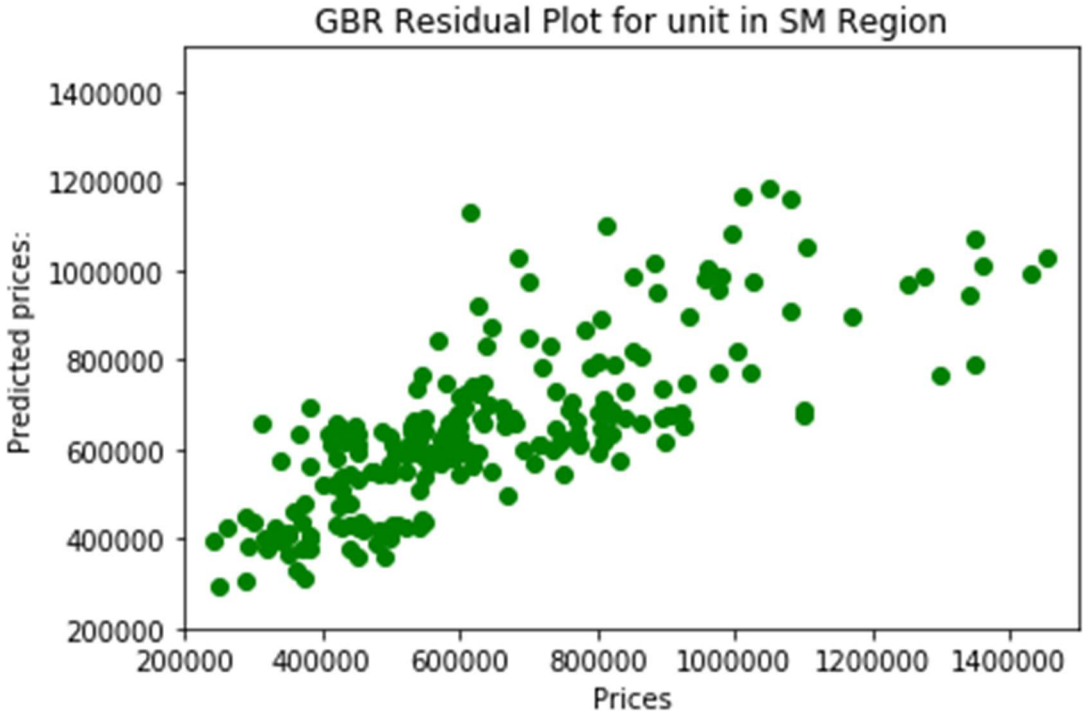
In this case Ridge regression was performed utilizing data frame df_S_M_u, which resulted in regression coefficient (R^2) of 0.64 and mean squared error (MSE) of 2.95E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 0 to optimize R^2 value.

Lasso Regression for type = unit in S M Region

In this case Lasso regression was performed utilizing data frame df_S_M_u, which resulted in regression coefficient (R^2) of 0.38 and mean squared error (MSE) of 4.92E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 1 to optimize R^2 value.

Gradient Booster Regressor for type = unit in S M Region

In this case gradient booster regression was performed utilizing data frame df_S_M_u, which resulted in regression coefficient (R^2) of 0.65 and mean squared error (MSE) of 2.86E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 1 and learning rate of 0.15 to optimize R^2 value.

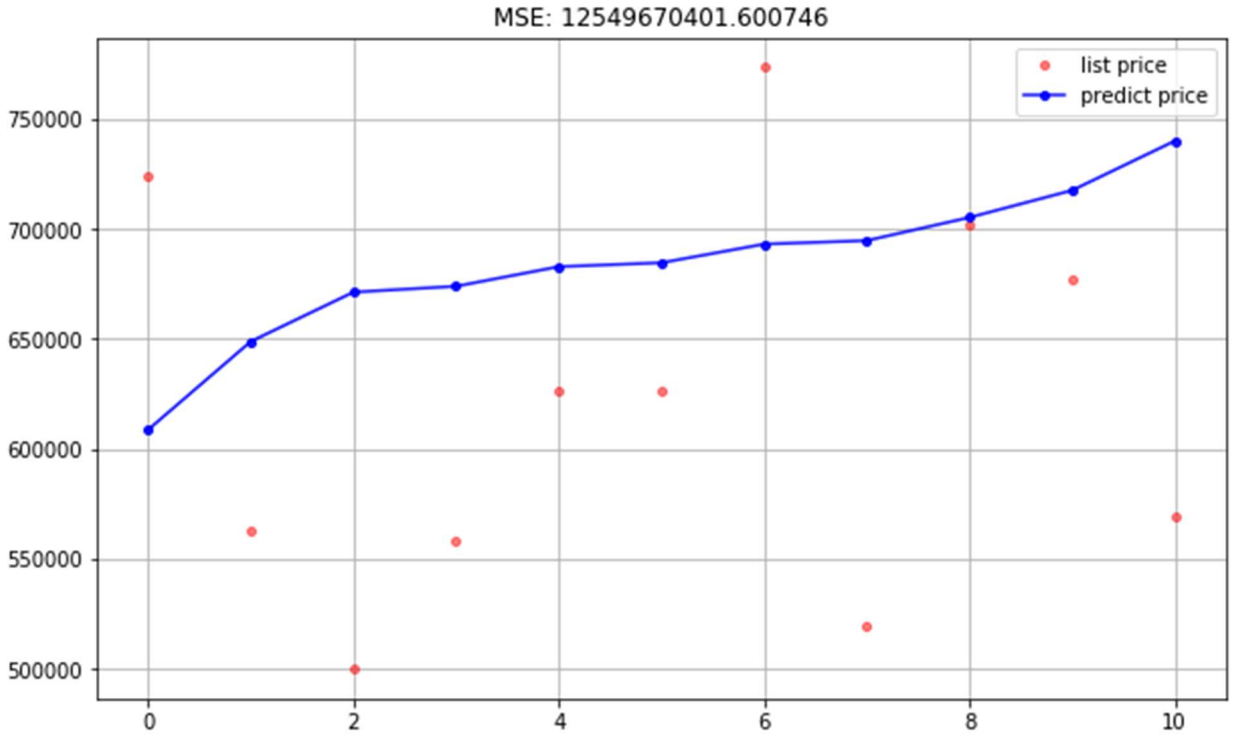


Summary of all regressions for type = unit in Southern Metropolitan Region is tabulated in table below;

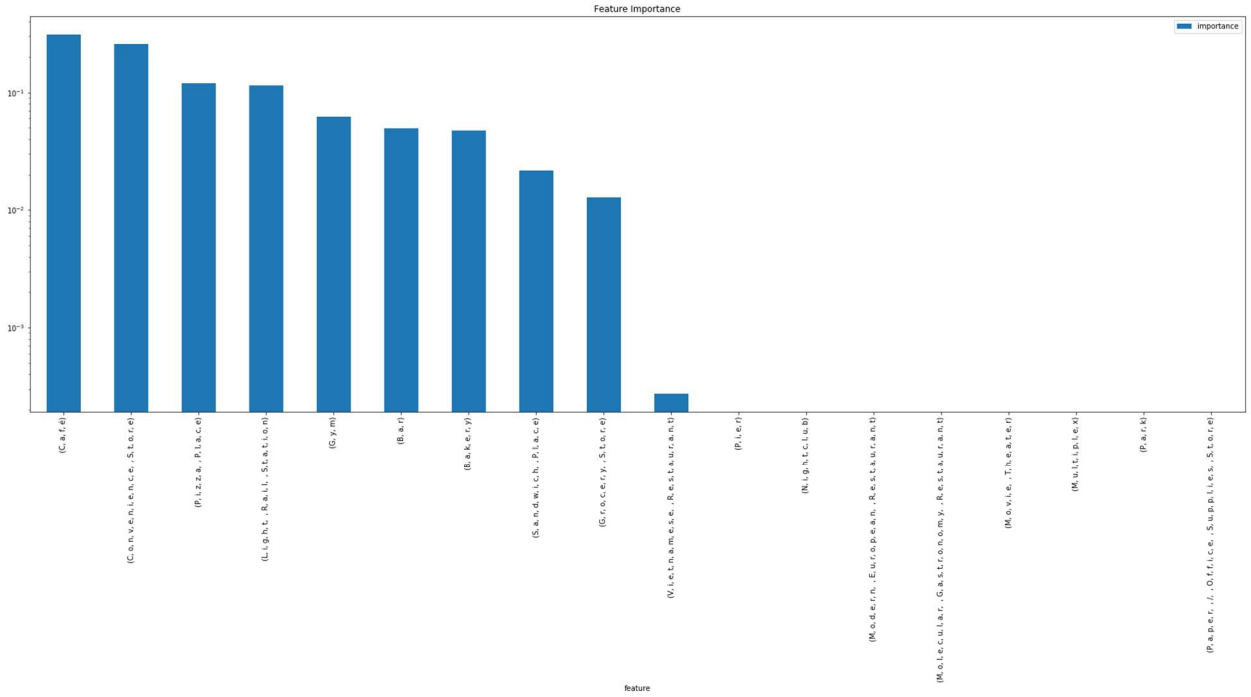
	Regression type	R^2	MAE	MSE	RMSE
Unit type in SM Region	Linear	0.36	141024	4.74E+10	217751
		0	190323	7.45E+10	273035
		0.25	167457	5.6E+10	236562
		0.01	189352	7.41E+10	272278
	MLR	0.49	129300	3.8E+10	194997
	Polynomial Pipe	0.53	123672	3.48E+10	186467
	Ridge	0.64	128885	2.95E+10	171897
	Lasso	0.38	140583	4.92E+10	221771
	GBR	0.65	126782	2.86E+10	169070

Random Forest Regression for type = unit in S M Region

Random Forest Regression was carried out to determine relationship between venue category and average property price. In this regard data frame joined_venues_price_S_M_u_df was utilized which resulted in regression coefficient (R^2) of ~ -0.75 with mean squared error (mse) of ~ 1.2E+10. Grid Search CV was utilized to determine best parameters including max_depth, min_samples_leaf & n_estimators in order to optimize R^2 value. Below plot shows list price versus predicted price;



In addition to this, feature importance using random forest regression was performed to determine venues with respect to their importance on average property prices of neighbourhoods. Below is the plot showing venue with its importance with respect to average property price;



Modeling for type = townhouse in Southern Metropolitan (S M) Region

Linear Regression

I. Number of rooms versus price for type = town in S M Region

In this case Linear regression was performed utilizing data frame df_S_M_t, which resulted in regression coefficient (R^2) of 0.18 and mean squared error (MSE) of 1.6E+11. Price of townhouse with 2 rooms was predicted, which found to be ~ $ 901739.

II. Year versus price for type = town in S M Region

In this case Linear regression was performed utilizing data frame df_S_M_t, which resulted in regression coefficient (R^2) of 0 and mean squared error (MSE) of 1.92E+11. Price of townhouse in year 2019 was predicted, which found to be ~ $ 1456648.

III. Distance from Central build up area (CBD) versus price for type = town in S M Region

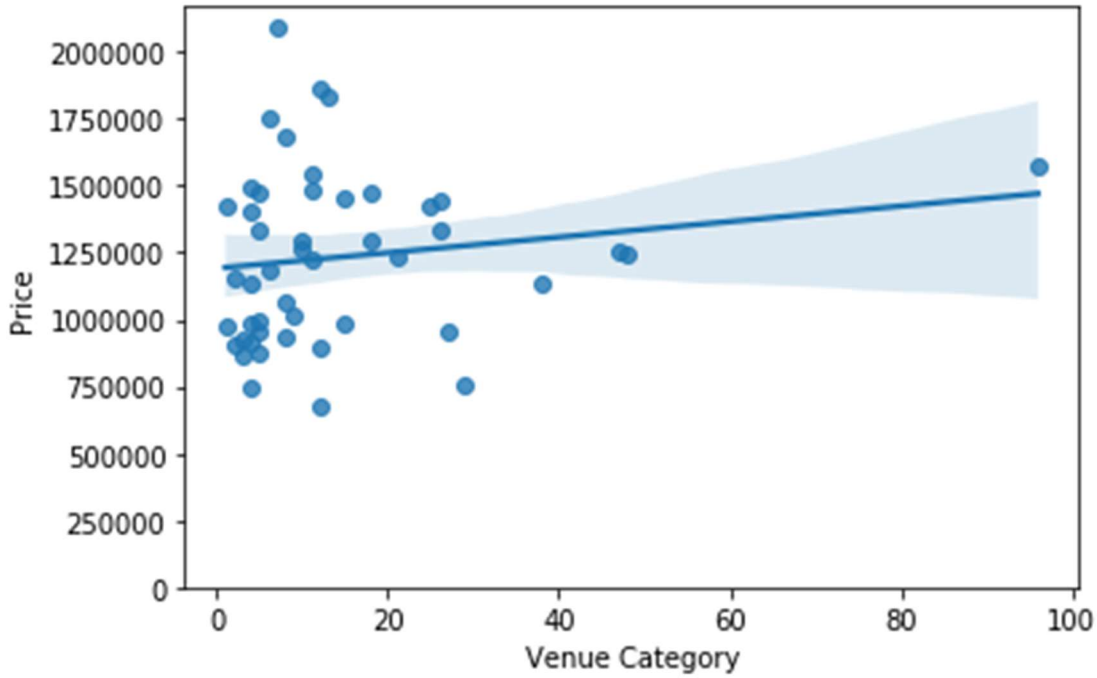
In this case Linear regression was performed utilizing data frame df_S_M_t, which resulted in regression coefficient (R^2) of 0.06 and mean squared error (MSE) of 1.85E+11. Price of townhouse 20 km from CBD was predicted, which found to be ~ $ 925264.

IV. Number of Bathroom versus price for type = town in S M Region

In this case Linear regression was performed utilizing data frame df_S_M_t, which resulted in regression coefficient (R^2) of 0.16 and mean squared error (MSE) of 1.64E+11. Price of townhouse with 2 bathrooms was predicted, which found to be ~ $ 1200950.

V. Venue category count versus price for type = town in S M Region

In this case seaborn plot was generated to investigate if linear relationship exists between venue count and average property price. Seaborn suggested poor relationship between venue count & average property price which was supplemented by linear regression model which computed regression coefficient(R^2) of 0.02 & mean squared error of 9.7E+10.



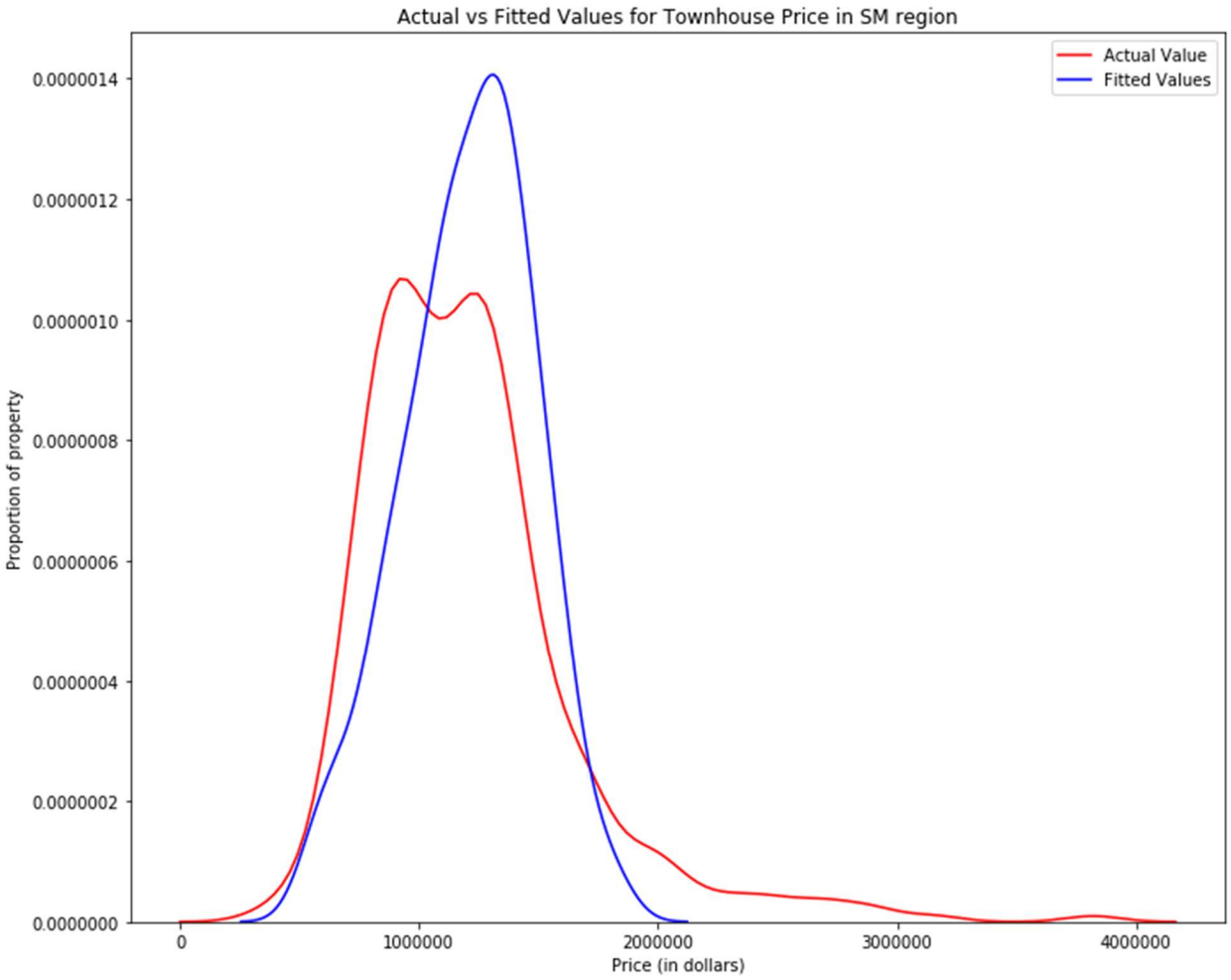
VI. Neighborhood venue category versus average property price for type = town in S M Region

In order to establish relationship between neighbourhood venue category versus average property price, v enue category from data frame south_metropolitan_t_venues were converted into binary format 0 & 1 using one hot encoding which was defined as a new data frame by the name of south_metropolitan_t_onehot. Neig hbourhood from south_metropolitan_t _venues was inserted into data frame south_metropolitan_t_onehot. Ve nue category in south_metropolitan_t_onehot was grouped & summed based on Neighbourhood and was def ined to a new data frame joined_venues_price_S_M_t_df. which was utilized for regression modelling which resulted in regression coefficient (R^2) of ~ -0.67 and mean squared error (mse) of 8.9E10.

MultiLinear Regression (MLR) for type = town in S M Region

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df_S_M_t, which resulted in regression coefficient (R^2) of 0.37 and mean squared error (MSE) of 1.229E+11. Price of townhouse with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 672380.

We can see that the fitted values are not reasonably close to the actual values, since the two distributions do not overlap.



Polynomial Regression with Pipe for type = town in S M Region

In this case polynomial pipe regression was performed utilizing data frame df_S_M_t, which resulted in regression coefficient (R^2) of 0.48 and mean squared error (MSE) of 1.027E+11. Price of townhouse with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 681997.

Ridge Regression for type = town in S M Region

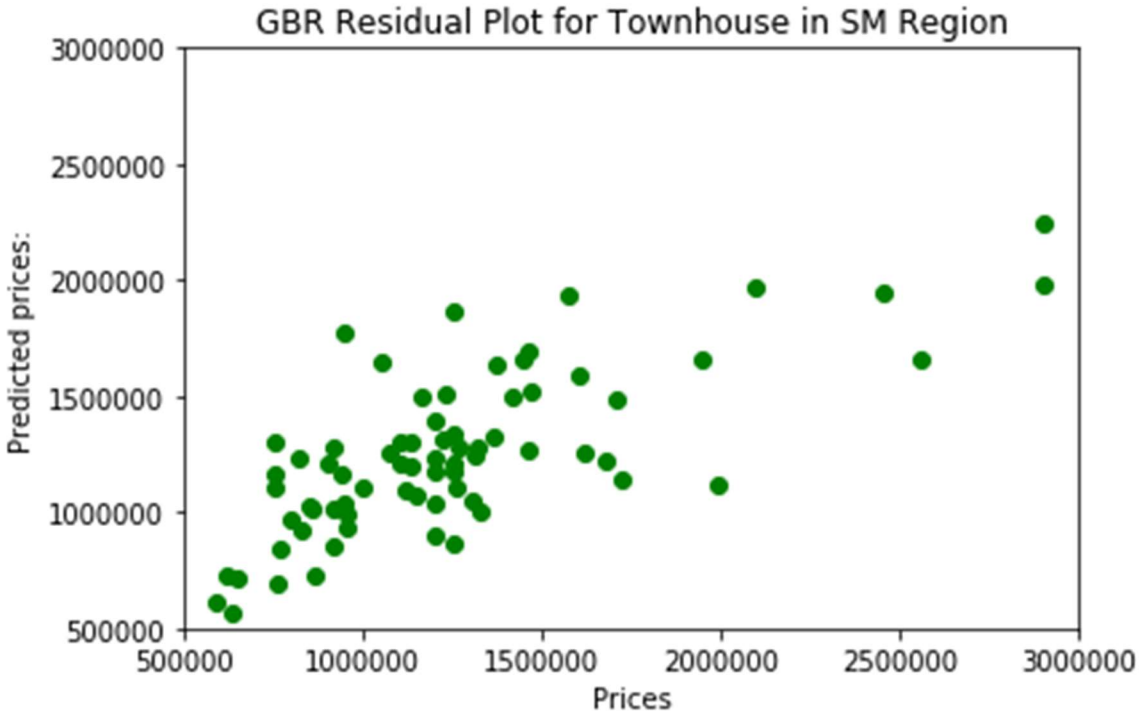
In this case Ridge regression was performed utilizing data frame df_S_M_t, which resulted in regression coefficient (R^2) of 0.44 and mean squared error (MSE) of 1.28E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 1 to optimize R^2 value.

Lasso Regression for type = town in S M Region

In this case Lasso regression was performed utilizing data frame df_S_M_t, which resulted in regression coefficient (R^2) of 0.35 and mean squared error (MSE) of 1.5E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 1 to optimize R^2 value.

Gradient Booster Regressor for type = town in S M Region

In this case gradient booster regression was performed utilizing data frame df_S_M_t, which resulted in regression coefficient (R^2) of 0.53 and mean squared error (MSE) of 1.08E+11. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 1 and learning rate of 0.15 to optimize R^2 value.

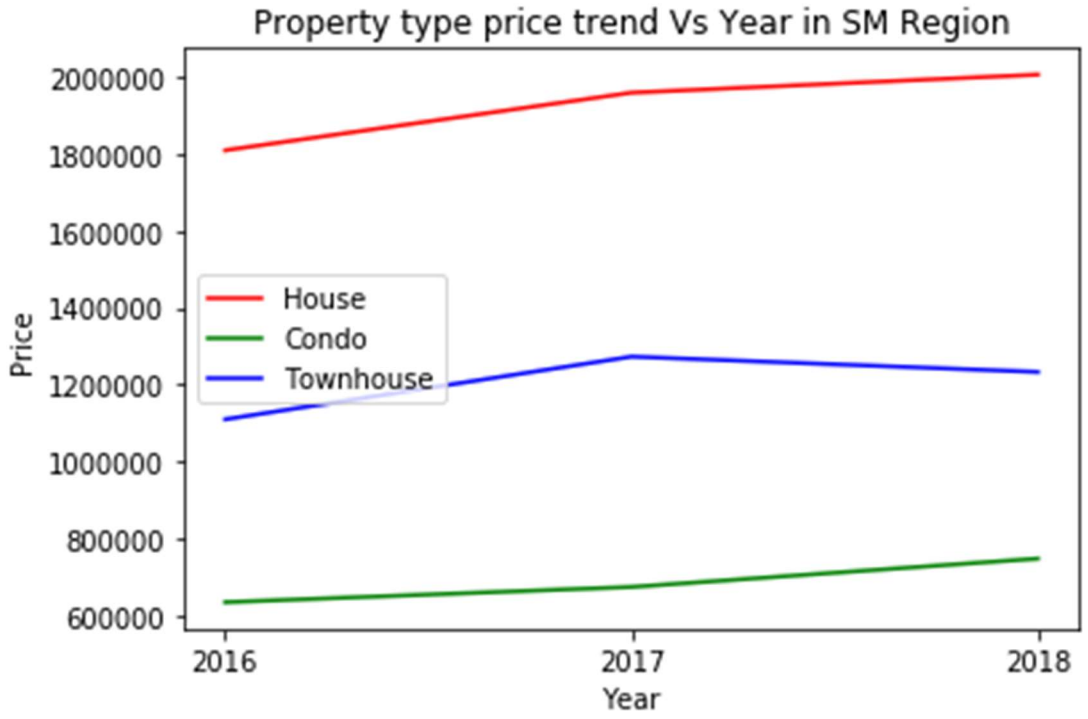


Summary of all regressions for type = town in Southern Metropolitan Region is tabulated in table below;

Townhouse type in S M Region	Regression type	R^2	MAE	MSE	RMSE
	Linear	0.18	283960	1.61282E+11	401599
		0.06	300037	1.85158E+11	430300
		0.16	287498	1.64088E+11	405077
		0	310288	1.92083E+11	438272
	MLR	0.37	238117	1.229E+11	350571
	Polynomial Pipe	0.48	217177	1.0275E+11	320546
	Ridge	0.44	248795	1.28728E+11	358786
	Lasso	0.35	268272	1.50724E+11	388232
	GBR	0.53	239904	1.08284E+11	329064

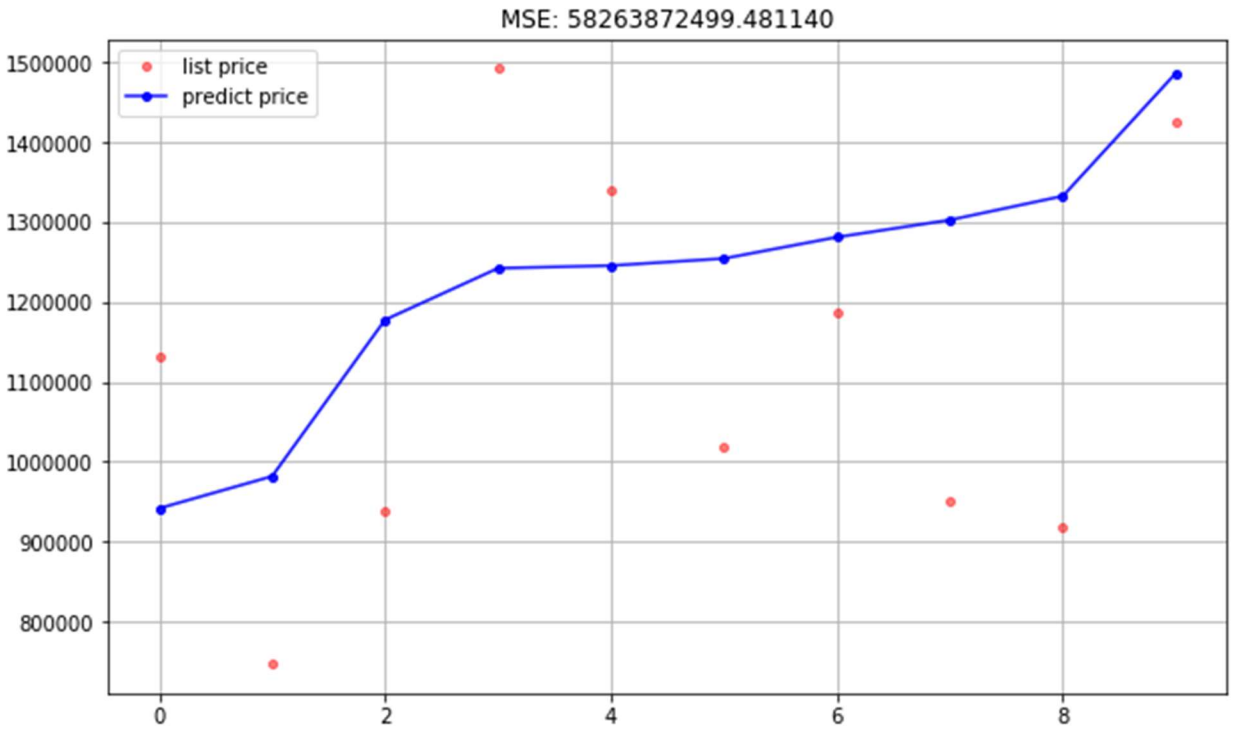
Line plot between mean price of each type in Southern Metropolitan

Graph suggests that House price increased by ~ $100,000/year, Condo price climbed up slowly while Townhouse price increased from 2016 to 2017 & then decreased from 2017 to 2018 in Southern Metropolitan region. It is time to built more condos & townhouses in 2019 due to minimal change in price over 3 year period. Furthermore, it can be concluded that sellers should be interested in selling Houses in coming years due to dramatic change in price.

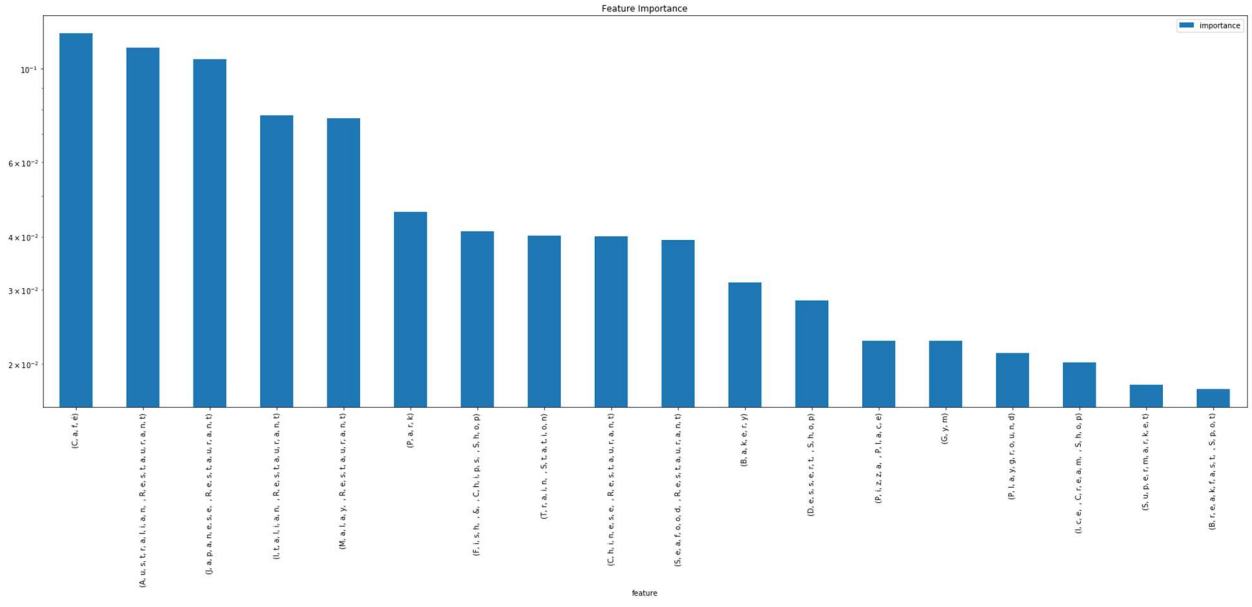


Random Forest Regression for type = town in S M Region

Random Forest Regression was carried out to determine relationship between venue category and average property price. In this regard data frame joined_venues_price_S_M_t_df was utilized which resulted in regression coefficient (R^2) of ~ -0.09 with mean squared error (mse) of ~ 5.8E+10. Grid Search CV was utilized to determine best parameters including max_depth, min_samples_leaf & n_estimators in order to optimize R^2 value. Below plot shows list price versus predicted price;



In addition to this, feature importance using random forest regression was performed to determine venues with respect to their importance on average property prices of neighbourhoods. Below is the plot showing venue with its importance with respect to average property price;



Northern Metropolitan (N M) Region

Linear Regression

I. Number of rooms versus price for all property types in N M Region

In this case Linear regression was performed utilizing data frame df_N_M, which resulted in regression coefficient (R^2) of 0.13 and mean squared error (MSE) of 1.71E+11. Price of property with 2 rooms was predicted, which found to be ~ $ 725713.

II. Year versus price for all property types in N M Region

In this case Linear regression was performed utilizing data frame df_N_M, which resulted in regression coefficient (R^2) of 0 and mean squared error (MSE) of 1.95E+11. Price of property in year 2019 was predicted, which found to be ~ $ 866567.

III. Distance from Central build up area (CBD) versus price for all property types in N M Region

In this case Linear regression was performed utilizing data frame df_N_M, which resulted in regression coefficient (R^2) of 0.16 and mean squared error (MSE) of 1.64E+11. Price of property 20 km from CBD was predicted, which found to be ~ $ 570957.

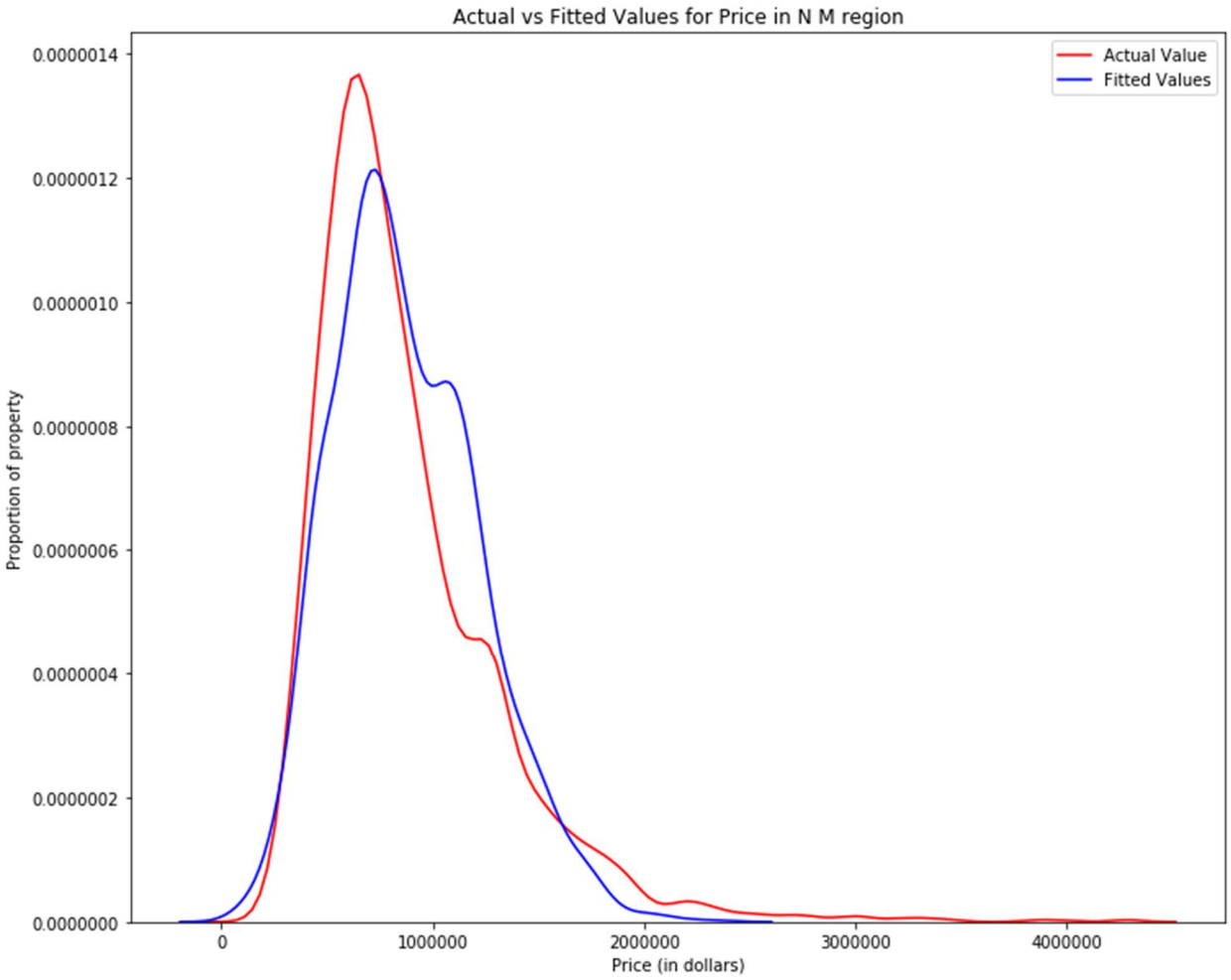
IV. Number of Bathroom versus price for all property types in N M Region

In this case Linear regression was performed utilizing data frame df_N_M, which resulted in regression coefficient (R^2) of 0.07 and mean squared error (MSE) of 1.825E+11. Price of property with 2 bathrooms was predicted, which found to be ~ $ 975305.

MultiLinear Regression (MLR) for all property types in N M Region

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df_N_M, which resulted in regression coefficient (R^2) of 0.61 and mean squared error (MSE) of 7.71E+10. Price of property with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 331443.

We can see that the fitted values are reasonably close to the actual values, since the two distributions overlap a bit. However, there is definitely some room for improvement.



Polynomial Regression with Pipe for all property types in N M Region

In this case polynomial pipe regression was performed utilizing data frame df_N_M, which resulted in regression coefficient (R^2) of 0.69 and mean squared error (MSE) of 6E+10. Price of property with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 492378.

Ridge Regression for all property types in N M Region

In this case Ridge regression was performed utilizing data frame df_N_M, which resulted in regression coefficient (R^2) of 0.7 and mean squared error (MSE) of 4.5E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 to optimize R^2 value.

Lasso Regression for all property types in N M Region

In this case Lasso regression was performed utilizing data frame df_N_M, which resulted in regression coefficient (R^2) of 0.61 and mean squared error (MSE) of 7.26E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 1 to optimize R^2 value.

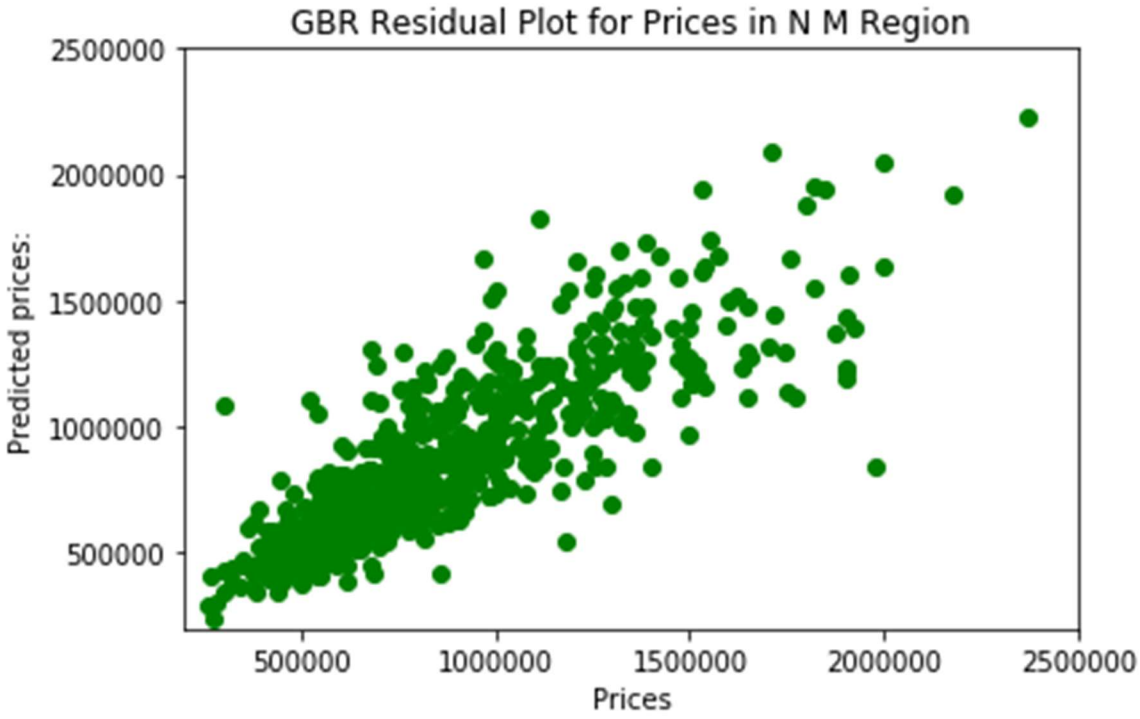
Principal Component Analysis

Principal component analysis (PCA), is a dimension reduction tool that projects data onto lower di- mensions, commonly referred to as principal components, in order to reduce the total number of variables to smaller data set with negligible information loss. In other words, if a feature is determined to be highly correlated to another, the feature is removed in order to help prevent overfitting of the model.

A heat map was constructed above confirming no notable correlation between the variables. In addition to this, a seaborn pairgrid plot was constructed to help visualise relationships between variables.

Gradient Booster Regressor

In this case gradient booster regression was performed utilizing data frame df_N_M, which resulted in regression coefficient (R^2) of 0.72 and mean squared error (MSE) of 4.18E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 and learning rate of 0.19 to optimize R^2 value.



Summary of all regressions for all types of property in Northern Metropolitan Region is tabulated in table below;

	Regression type	R^2	MAE	MSE	RMSE
All types in Northern Metropolitan Region	Linear	0.13	307566	1.71E+11	413471
		0.16	280010	1.64E+11	405030
		0.07	318820	1.825E+11	427244
		0	321918	1.957E+11	442327
	MLR	0.61	191768	7.717E+10	277798
	Polynomial Pipe	0.69	167021	6E+10	244950
	Ridge	0.7	151614	4.511E+10	212391
	Lasso	0.61	195388	7.266E+10	269551
	PCA	0.72	141530	4.181E+10	204465

Modeling for type = house in Northern Metropolitan (N M) Region

Linear Regression

I. Number of rooms versus price for type = house in N M Region

In this case Linear regression was performed utilizing data frame df_N_M_h, which resulted in regression coefficient (R^2) of 0.02 and mean squared error (MSE) of 2.03E+11. Price of house with 2 rooms was predicted, which found to be ~ $ 884373.

II. Year versus price for type = house in N M Region

In this case Linear regression was performed utilizing data frame df_N_M_h, which resulted in regression coefficient (R^2) of 0.02 and mean squared error (MSE) of 2.049E+11. Price of house in year 2019 was predicted, which found to be ~ $ 795834.

III. Distance from Central build up area (CBD) versus price for type = house in N M Region

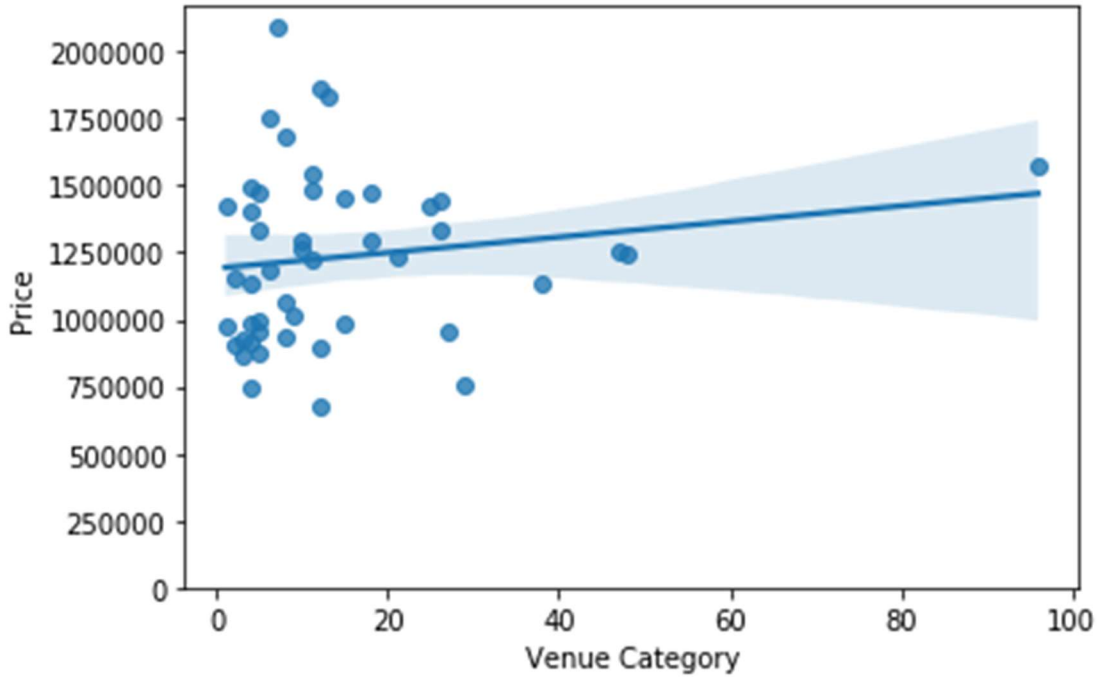
In this case Linear regression was performed utilizing data frame df_N_M_h, which resulted in regression coefficient (R^2) of 0.39 and mean squared error (MSE) of 1.27E+11. Price of house 20 km from CBD was predicted, which found to be ~ $ 557449.

IV. Number of Bathroom versus price for type = house in N M Region

In this case Linear regression was performed utilizing data frame df_n_M_h, which resulted in regression coefficient (R^2) of 0.02 and mean squared error (MSE) of 2.049+11. Price of house with 2 bathrooms was predicted, which found to be ~ $ 1036375.

V. Venue category count versus price for type = house in N M Region

In this case seaborn plot was generated to investigate if linear relationship exists between venue count and average property price. Seaborn suggested poor relationship between venue count & average property price which was supplemented by linear regression model which computed regression coefficient(R^2) of 0.13 & mean squared error of 316028.



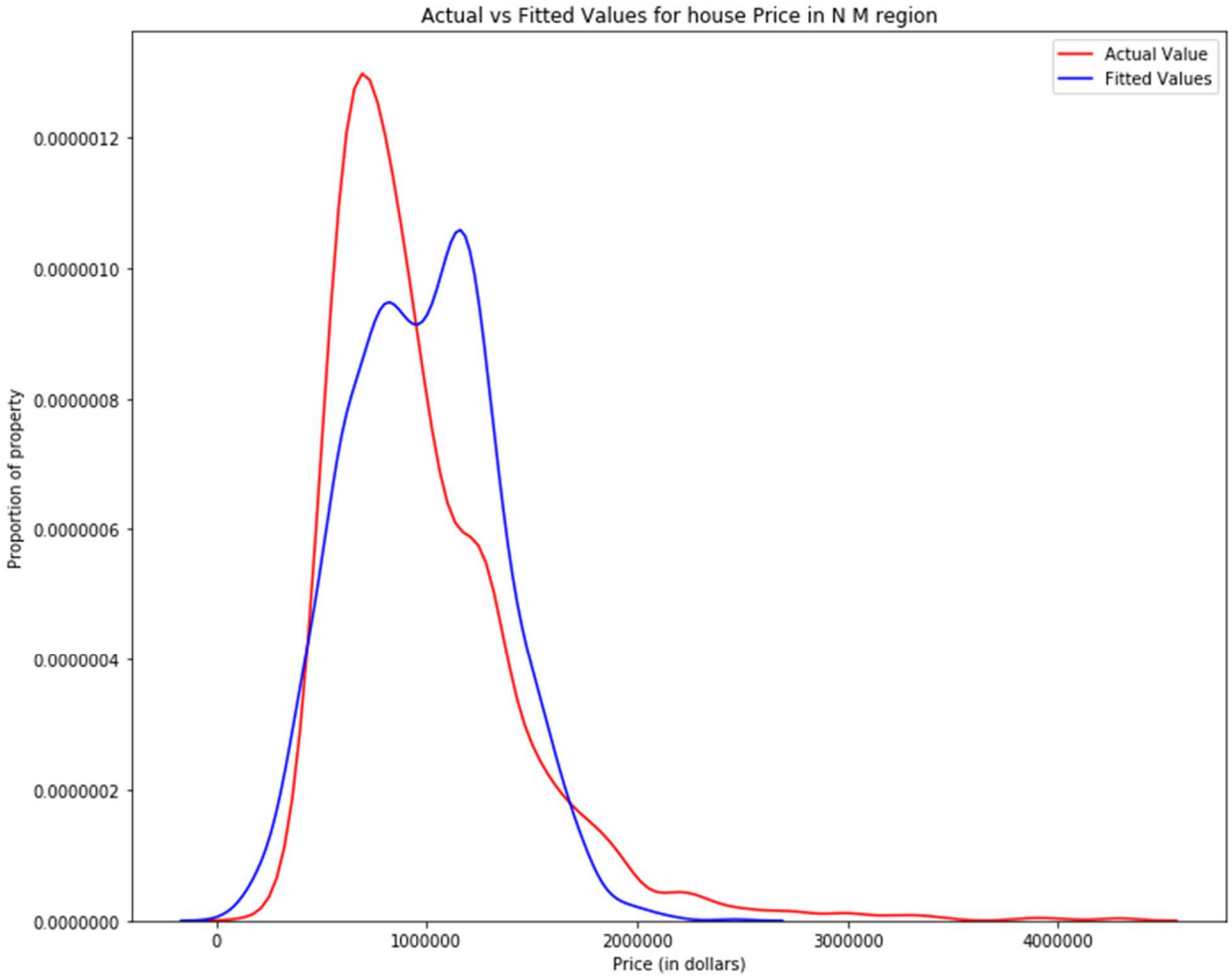
VI. Neighborhood venue category versus average property price for type = house in N M Region

In order to establish relationship between neighbourhood venue category versus average property price, v enue category from data frame north_metropolitan_h_venues were converted into binary format 0 & 1 using one hot encoding which was defined as a new data frame by the name of north_metropolitan_h_onehot. Nei ghbourhood from north_metropolitan_h _venues was inserted into data frame north_metropolitan_h_onehot. Venue category in north_metropolitan_h_onehot was grouped & summed based on Neighbourhood and was defined to a new data frame joined_venues_price_N_M_h_df. which was utilized for regression modelling w hich resulted in regression coefficient (R^2) of ~ 0.405 and mean squared error (mse) of 5.27E10.

MultiLinear Regression (MLR) for type = house in N M Region

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df_N_M_h, which resulted in regression coefficient (R^2) of 0.57 and mean squared error (MSE) of 8.5E+10. Price of house with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 382061.

We can see that the fitted values are not reasonably close to the actual values, since the two distributions do not overlap a bit.



Polynomial Regression with Pipe for type = house in N M Region

In this case polynomial pipe regression was performed utilizing data frame df_N_M_h, which resulted in regression coefficient (R^2) of 0.7 and mean squared error (MSE) of 6.32E+10. Price of house with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 400445.

Ridge Regression for type = house in N M Region

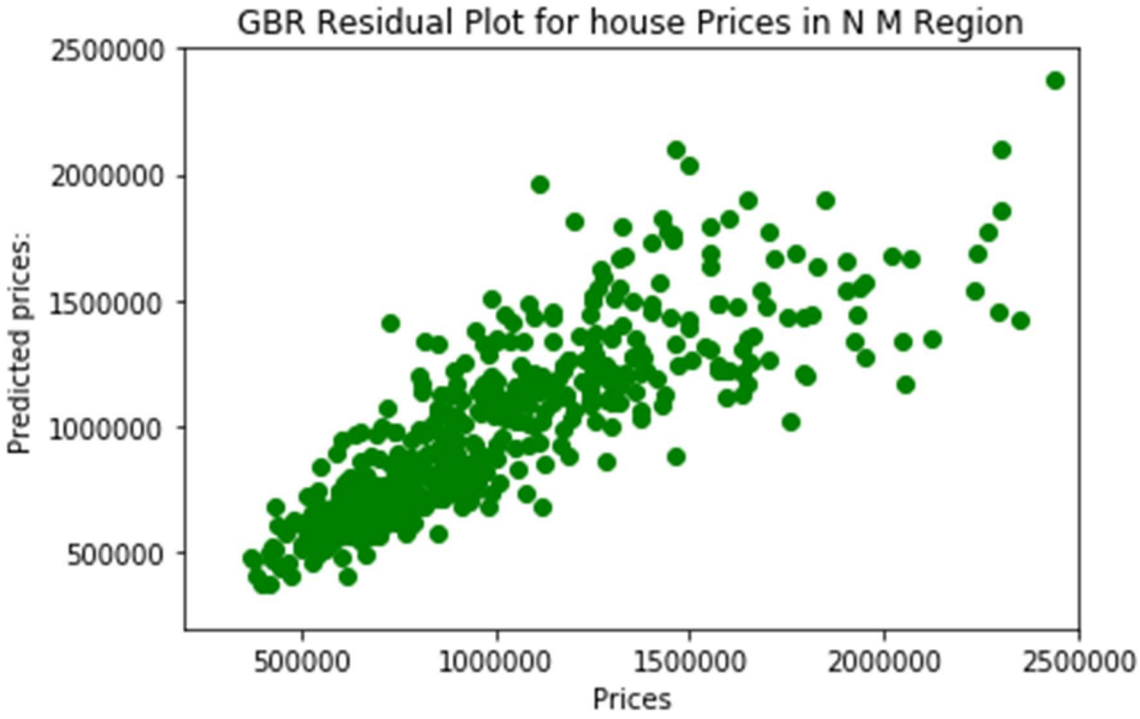
In this case Ridge regression was performed utilizing data frame df_N_M_h, which resulted in regression coefficient (R^2) of 0.65 and mean squared error (MSE) of 7.76E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 0 to optimize R^2 value.

Lasso Regression for type = house in N M Region

In this case Lasso regression was performed utilizing data frame df_N_M_h, which resulted in regression coefficient (R^2) of 0.61 and mean squared error (MSE) of 7.43E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 3 to optimize R^2 value.

Gradient Booster Regressor for type = house in N M Region

In this case gradient booster regression was performed utilizing data frame df_N_M_h, which resulted in regression coefficient (R^2) of 0.7 and mean squared error (MSE) of 6.57E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 0 and learning rate of 0.19 to optimize R^2 value.

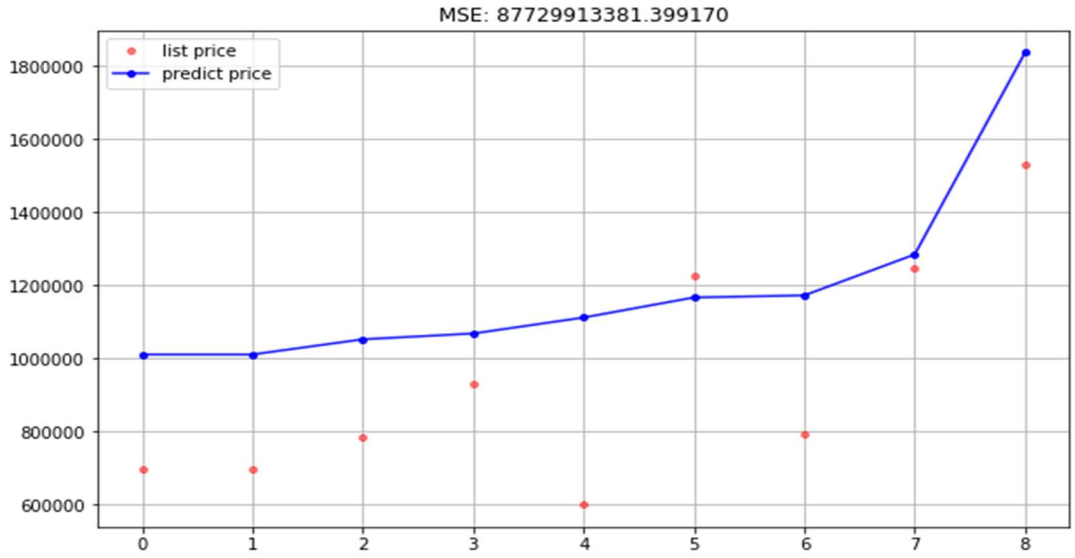


Summary of all regressions for type = house in Northern Metropolitan Region is tabulated in table below;

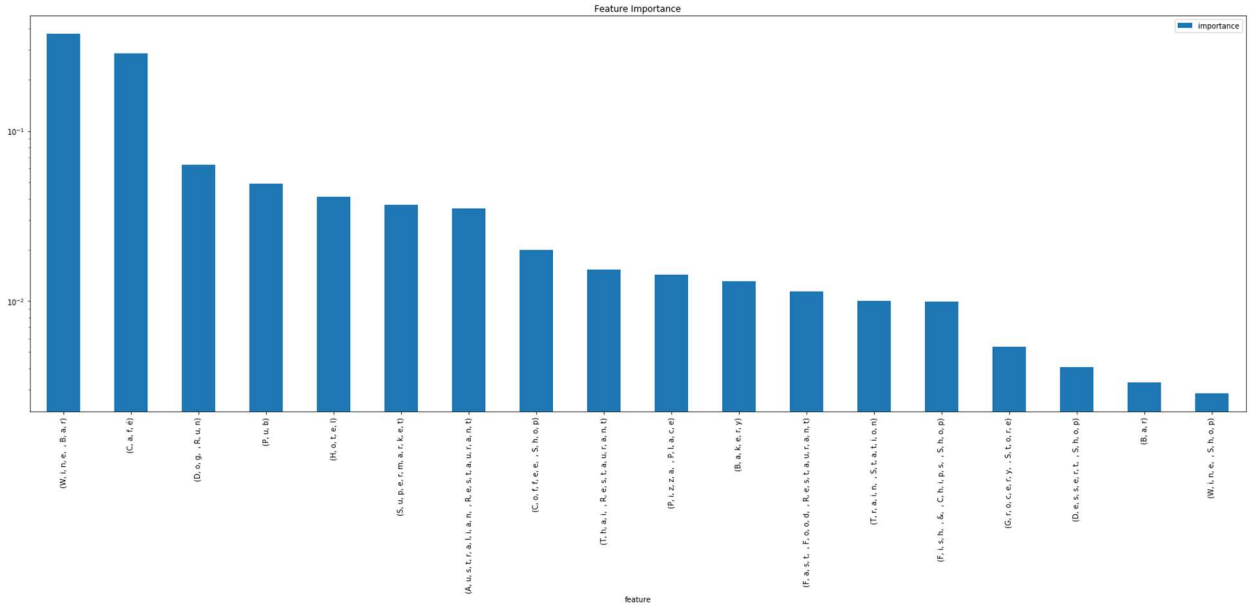
	Regression type	R^2	MAE	MSE	RMSE
House Type Northern Metropolitan Region	Linear	0.02	337502	2.0372E+11	451358
		0.39	239355	1.27E+11	356370
		0.03	334647	2.0248E+11	449980
		0.02	329994	2.0491E+11	452669
	MLR	0.57	196862	8.5058E+10	291646
	Polynomial Pipe	0.7	167918	6.3257E+10	251509
	Ridge	0.65	182592	7.7665E+10	278683
	Lasso	0.61	195348	7.4304E+10	272588
	GBR	0.7	167309	6.5795E+10	256504

Random Forest Regression for type = house in N M Region

Random Forest Regression was carried out to determine relationship between venue category and average property price. In this regard data frame joined_venues_price_N_M_h_df was utilized which resulted in regression coefficient (R^2) of ~ -0.0127 with mean squared error (mse) of ~ 8.7E+10. Grid Search CV was utilized to determine best parameters including max_depth, min_samples_leaf & n_estimators in order to optimize R^2 value. Below plot shows list price versus predicted price;



In addition to this, feature importance using random forest regression was performed to determine venues with respect to their importance on average property prices of neighbourhoods. Below is the plot showing venue with its importance with respect to average property price;



Modeling for type = unit in Northern Metropolitan (N M) Region

Linear Regression

I. Number of rooms versus price for type = unit in N M Region

In this case Linear regression was performed utilizing data frame df_N_M_u, which resulted in regression coefficient (R^2) of 0.22 and mean squared error (MSE) of 3.52E+10. Price of unit with 2 rooms was predicted, which found to be ~ $ 559261.

II. Year versus price for type = unit in N M Region

In this case Linear regression was performed utilizing data frame df_N_M_u, which resulted in regression coefficient (R^2) of 0.01 and mean squared error (MSE) of 4.47E+10. Price of unit in year 2019 was predicted, which found to be ~ $ 611403.

III. Distance from Central build up area (CBD) versus price for type = unit in N M Region

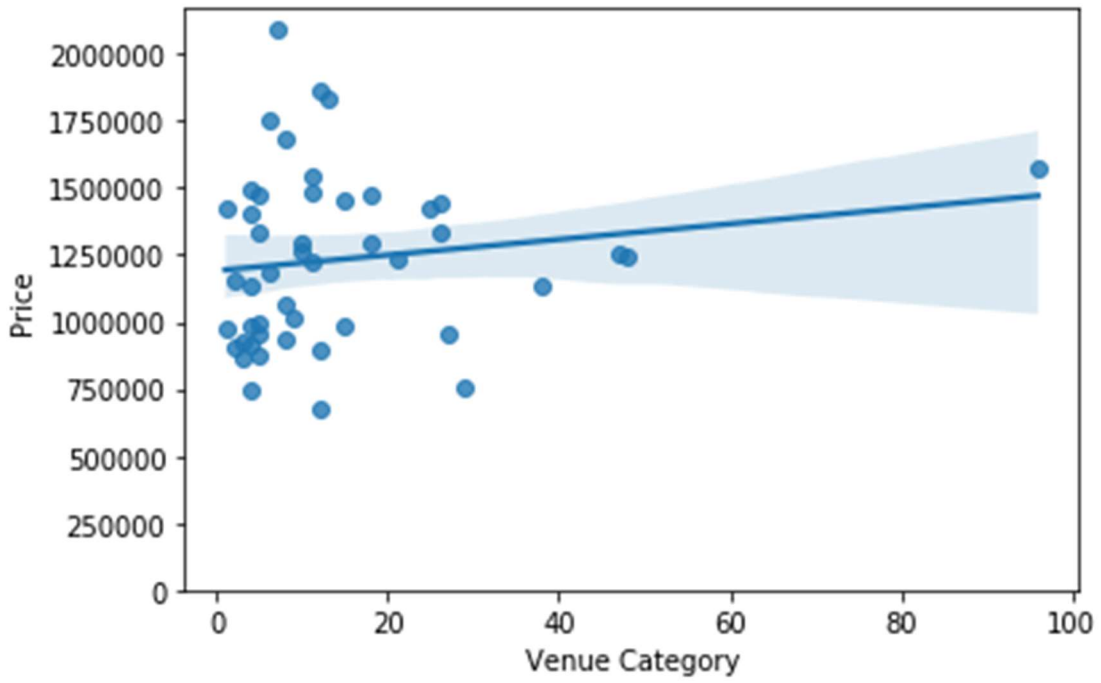
In this case Linear regression was performed utilizing data frame df_N_M_u, which resulted in regression coefficient (R^2) of 0.08 and mean squared error (MSE) of 4.13E+10. Price of unit 20 km from CBD was predicted, which found to be ~ $ 330682.

IV. Number of Bathroom versus price for type = unit in N M Region

In this case Linear regression was performed utilizing data frame df_N_M_u, which resulted in regression coefficient (R^2) of 0.24 and mean squared error (MSE) of 3.41E+10. Price of unit with 2 bathrooms was predicted, which found to be ~ $ 774839.

V. Venue category count versus price for type = unit in N M Region

In this case seaborn plot was generated to investigate if linear relationship exists between venue count and average property price. Seaborn suggested poor relationship between venue count & average property price which was supplemented by linear regression model which computed regression coefficient(R^2) of 0.12 & mean squared error of 1.5E10.



VI. Neighborhood venue category versus average property price for type = unit in N M Region

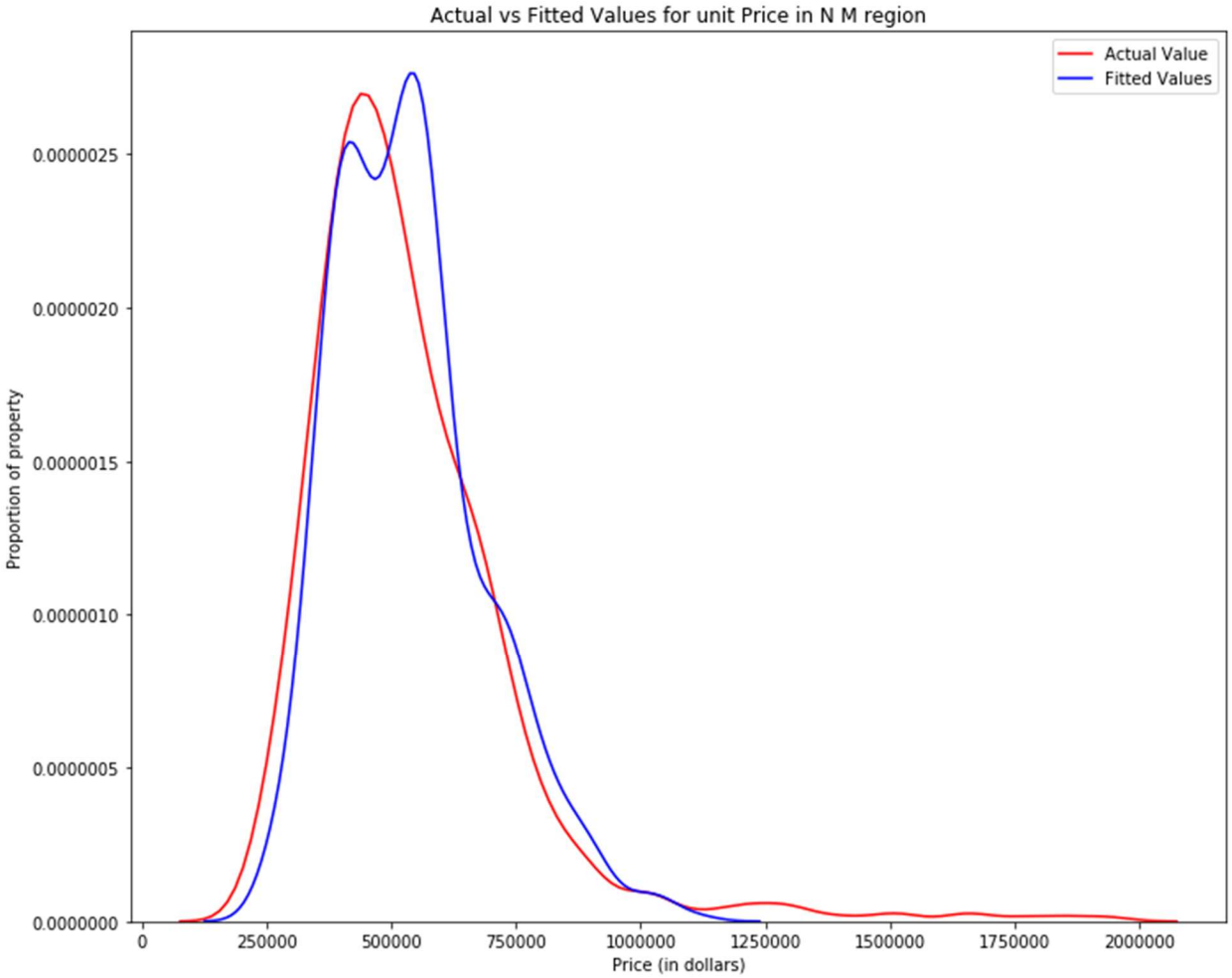
In order to establish relationship between neighbourhood venue category versus average property price, v enue category from data frame north_metropolitan_u_venues were converted into binary format 0 & 1 using one hot encoding which was defined as a new data frame by the name of north_metropolitan_u_onehot. Nei ghbourhood from north_metropolitan_u_venues was inserted into data frame north_metropolitan_u_onehot. V enue category in north_metropolitan_u_onehot was grouped & summed based on Neighbourhood and was d efined to a new data frame joined_venues_price_N_M_u_df. which was utilized for regression modelling wh ich resulted in regression coefficient (R^2) of -0.67 and mean squared error (mse) of 1.03E10.

MultiLinear Regression (MLR) for type = unit in N M Region

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df_N_M_u, which resulted in regression

coefficient (R^2) of 0.51 and mean squared error (MSE) of 2.21E+10. Price of unit with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 520936.

We can see that the fitted values are reasonably close to the actual values, since the two distributions overlap a bit but there needs an improvement.



Polynomial Regression with Pipe for type = unit in N M Region

In this case polynomial pipe regression was performed utilizing data frame df_N_M_u, which resulted in regression coefficient (R^2) of 0.63 and mean squared error (MSE) of 1.68E+10. Price of unit with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 425076.

Ridge Regression for type = unit in N M Region

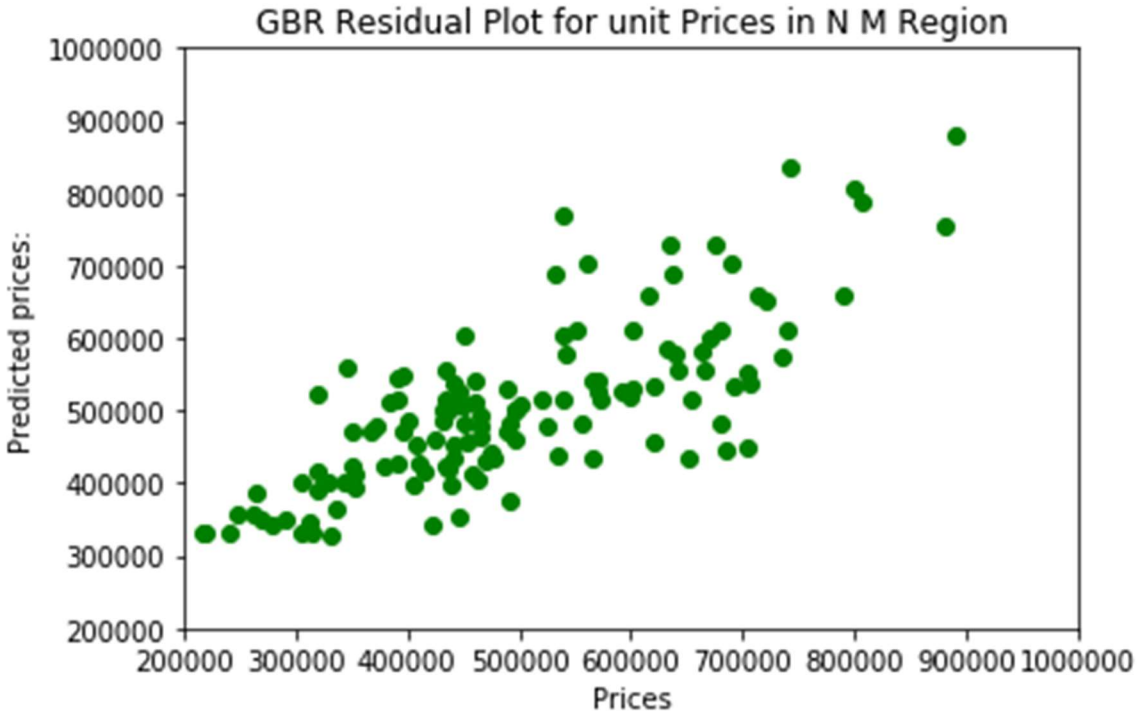
In this case Ridge regression was performed utilizing data frame df_N_M_u, which resulted in regression coefficient (R^2) of 0.6 and mean squared error (MSE) of 1.95E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 to optimize R^2 value.

Lasso Regression for type = unit in N M Region

In this case Lasso regression was performed utilizing data frame df_N_M_u, which resulted in regression coefficient (R^2) of 0.53 and mean squared error (MSE) of 1.88E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 4 to optimize R^2 value.

Gradient Booster Regressor for type = unit in N M Region

In this case gradient booster regression was performed utilizing data frame df_N_M_u, which resulted in regression coefficient (R^2) of 0.56 and mean squared error (MSE) of 2.16E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 and learning rate of 0.1 to optimize R^2 value.

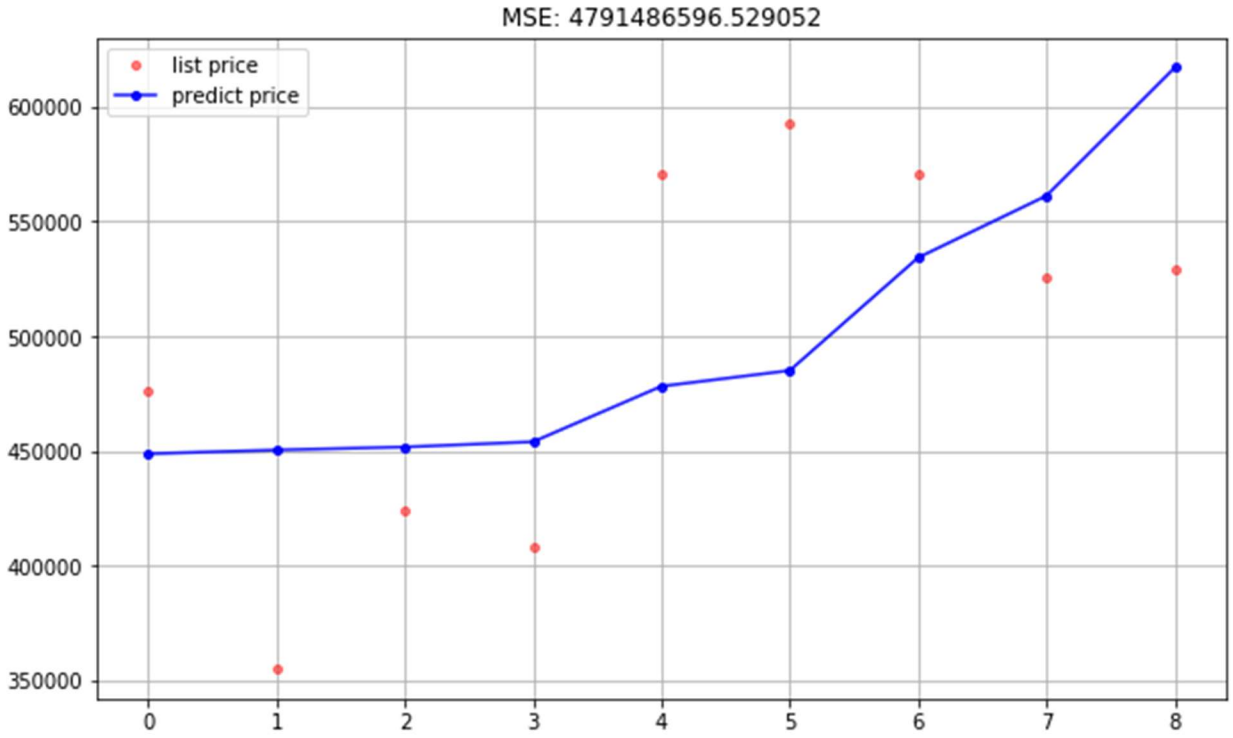


Summary of all regressions for type = unit in Northern Metropolitan Region is tabulated in table below;

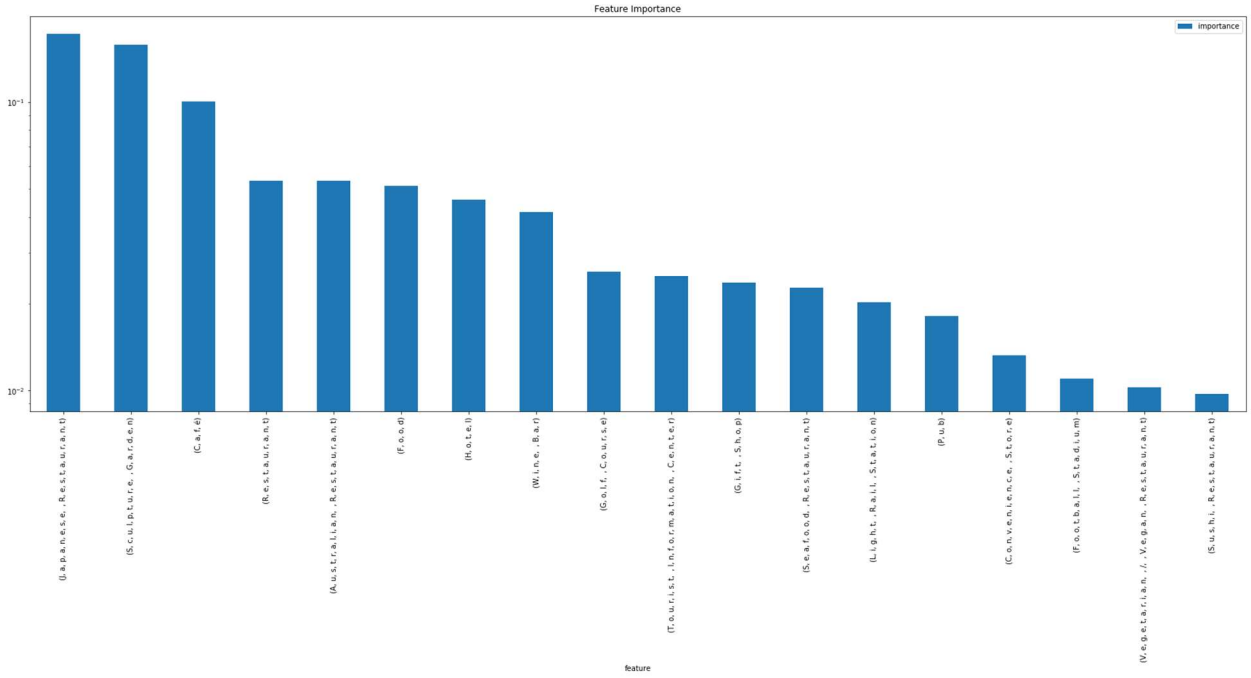
	Regression type	R^2	MAE	MSE	RMSE
Unit type in NM Region	Linear	0.22	124659	3.52E+10	187724
		0.08	139802	4.13E+10	203115
		0.24	126641	3.41E+10	184537
		0.01	144254	4.47E+10	211448
	MLR	0.51	100431	2.21E+10	148707
	Polynomial Pipe	0.63	90004	1.68E+10	129569
	Ridge	0.6	91011	1.95E+10	139656
	Lasso	0.53	94326	1.88E+10	137100
	GBR	0.56	92110	2.16E+10	146907

Random Forest Regression for type = unit in N M Region

Random Forest Regression was carried out to determine relationship between venue category and average property price. In this regard data frame joined_venues_price_N_M_u_df was utilized which resulted in regression coefficient (R^2) of ~ -0.22 with mean squared error (mse) of ~ 4.7E+9. Grid Search CV was utilized to determine best parameters including max_depth, min_samples_leaf & n_estimators in order to optimize R^2 value. Below plot shows list price versus predicted price;



In addition to this, feature importance using random forest regression was performed to determine venues with respect to their importance on average property prices of neighbourhoods. Below is the plot showing venue with its importance with respect to average property price;



Modeling for type = townhouse in Northern Metropolitan (N M) Region

Linear Regression

I. Number of rooms versus price for type = town in N M Region

In this case Linear regression was performed utilizing data frame df_N_M_t, which resulted in regression coefficient (R^2) of 0.04 and mean squared error (MSE) of 6.23E+10. Price of townhouse with 2 rooms was predicted, which found to be ~ $ 703287.

II. Year versus price for type = town in N M Region

In this case Linear regression was performed utilizing data frame df_N_M_t, which resulted in regression coefficient (R^2) of 0.01 and mean squared error (MSE) of 6.427E+10. Price of townhouse in year 2019 was predicted, which found to be ~ $ 819872.

III. Distance from Central build up area (CBD) versus price for type = town in N M Region

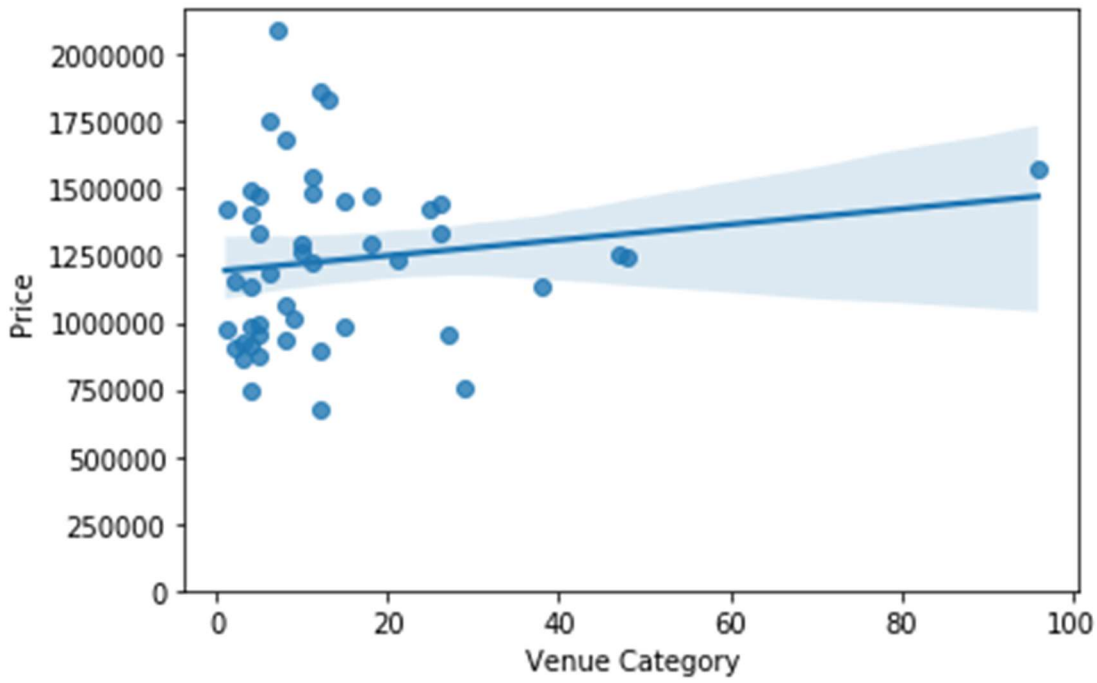
In this case Linear regression was performed utilizing data frame df_N_M_t, which resulted in regression coefficient (R^2) of 0.49 and mean squared error (MSE) of 3.3E+10. Price of townhouse 20 km from CBD was predicted, which found to be ~ $ 294654.

IV. Number of Bathroom versus price for type = town in N M Region

In this case Linear regression was performed utilizing data frame df_N_M_t, which resulted in regression coefficient (R^2) of 0.12 and mean squared error (MSE) of 5.66E+10. Price of townhouse with 2 bathrooms was predicted, which found to be ~ $ 811421.

V. Venue category count versus price for type = town in N M Region

In this case seaborn plot was generated to investigate if linear relationship exists between venue count and average property price. Seaborn suggested poor relationship between venue count & average property price which was supplemented by linear regression model which computed regression coefficient(R^2) of 0.37 & mean squared error of 4.47E10.



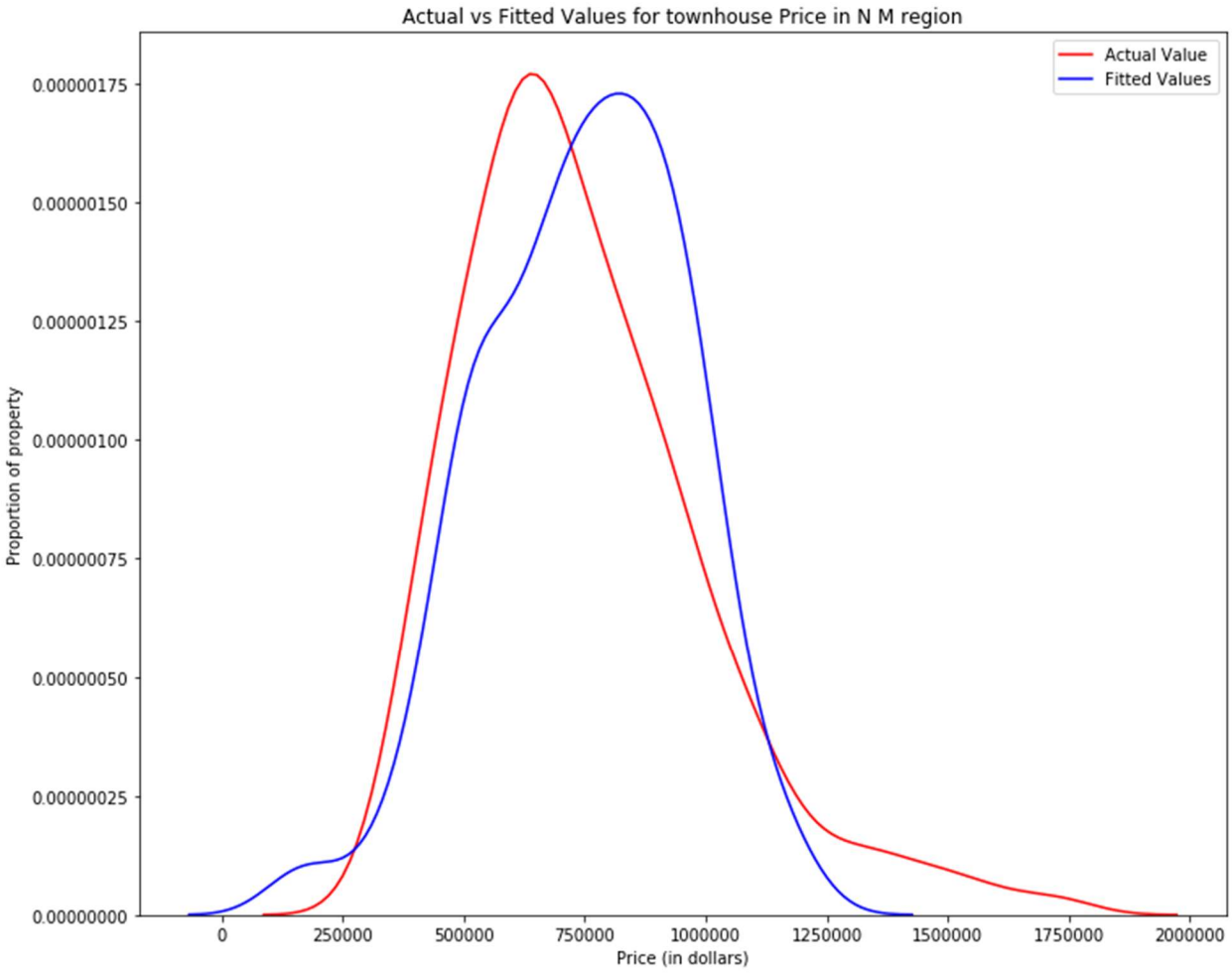
VI. Neighborhood venue category versus average property price for type = town in N M Region

In order to establish relationship between neighbourhood venue category versus average property price, v enue category from data frame north_metropolitan_t_venues were converted into binary format 0 & 1 using one hot encoding which was defined as a new data frame by the name of north_metropolitan_t_onehot. Neig hbourhood from north_metropolitan_t_venues was inserted into data frame north_metropolitan_t_onehot. Ven ue category in north_metropolitan_t_onehot was grouped & summed based on Neighbourhood and was defi ned to a new data frame joined_venues_price_N_M_t_df. which was utilized for regression modelling which resulted in regression coefficient (R^2) of 0.23 and mean squared error (mse) of 5.2E10.

MultiLinear Regression (MLR) for type = town in N M Region

Independent variables including Rooms, Distance, Bathroom, Year, Yearbuilt, Car & Landsize were used to see the relationship with price. Same variables will be used in Ridge, Lasso & PCA. In this case Multilinear regression was performed utilizing data frame df_N_M_t, which resulted in regression coefficient (R^2) of 0.66 and mean squared error (MSE) of 2.177E+10. Price of townhouse with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 411856.

We can see that the fitted values are not reasonably close to the actual values, since the two distributions do not overlap.



Polynomial Regression with Pipe for type = town in N M Region

In this case polynomial pipe regression was performed utilizing data frame df_N_M_t, which resulted in regression coefficient (R^2) of 0.75 and mean squared error (MSE) of 1.64E+10. Price of townhouse with 2 bathrooms, 2 rooms, 20 km away from CBD in year 2019 was predicted, which found to be ~ $ 559303.

Ridge Regression for type = town in N M Region

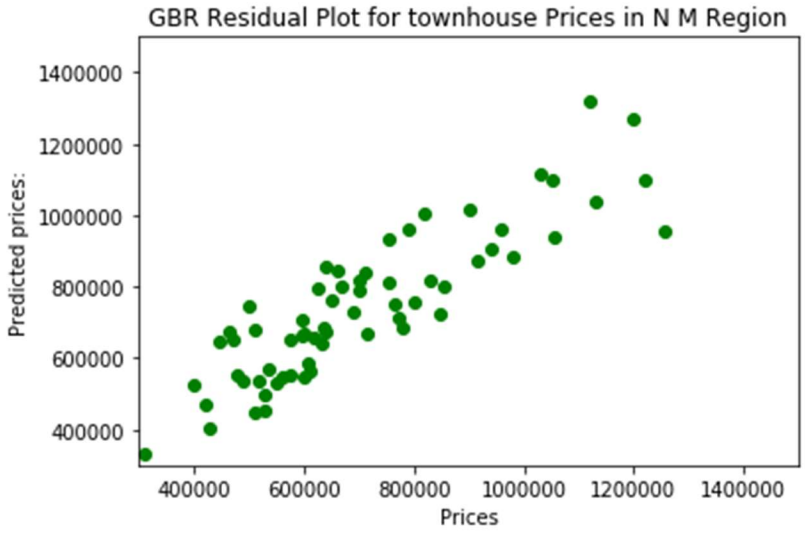
In this case Ridge regression was performed utilizing data frame df_N_M_t, which resulted in regression coefficient (R^2) of 0.61 and mean squared error (MSE) of 2.64E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 3 to optimize R^2 value.

Lasso Regression for type = town in N M Region

In this case Lasso regression was performed utilizing data frame df_N_M_t, which resulted in regression coefficient (R^2) of 0.58 and mean squared error (MSE) of 1.95E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 to optimize R^2 value.

Gradient Booster Regressor for type = town in N M Region

In this case gradient booster regression was performed utilizing data frame df_N_M_t, which resulted in regression coefficient (R^2) of 0.77 and mean squared error (MSE) of 1.25E+10. Train test split was used to split the entire data in to 20% for testing & 80% for training with random state of 2 and learning rate of 0.15 to optimize R^2 value.



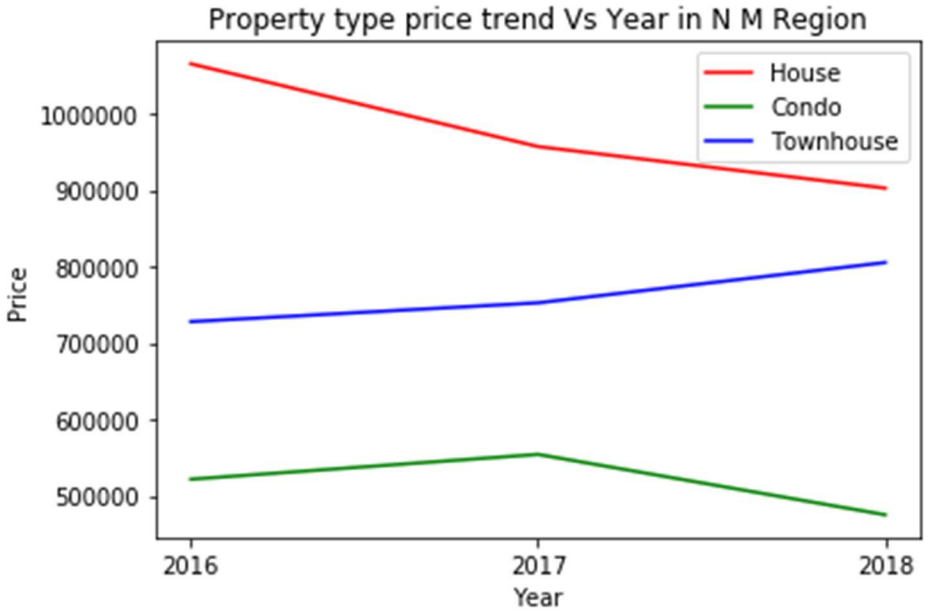
Summary of all regressions for type = town in Northern Metropolitan Region is tabulated in table below;

	Regression type	R^2	MAE	MSE	RMSE
Townhouse type in N M Region	Linear	0.04	192099	6.234E+10	249679
		0.49	132854	3.301E+10	181679
		0.12	183837	5.669E+10	238089
		0.01	196135	6.427E+10	253522
	MLR	0.66	109784	2.177E+10	147552
	Polynomial Pipe	0.75	94928	1.602E+10	126574
	Ridge	0.61	117792	2.646E+10	162673
	Lasso	0.58	115083	1.953E+10	139734
	GBR	0.77	89951	1.253E+10	111941

Line plot between mean price of all three types in Northern Metropolitan region.

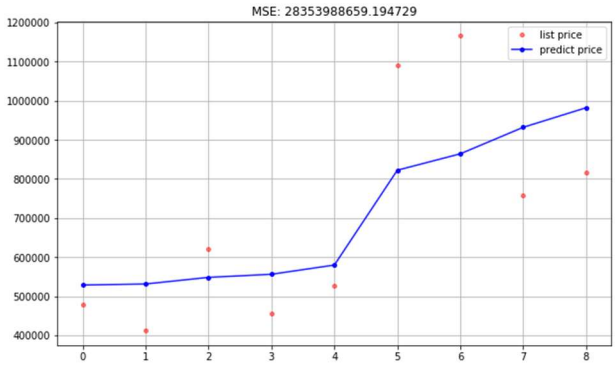
Graph suggests that House price decreased by ~ $50,000/year, Condo price climbed up slowly in 2017 but decreased in 2018 while Townhouse price increased by ~ 100,000 in 2 years in Northern Metropolitan region. It

is time to built more condos & houses in 2019 due to decrease in price over 2 year period. Furthermore, it can be concluded that sellers should be interested in selling Townhouses in coming years due to dramatic change in price.

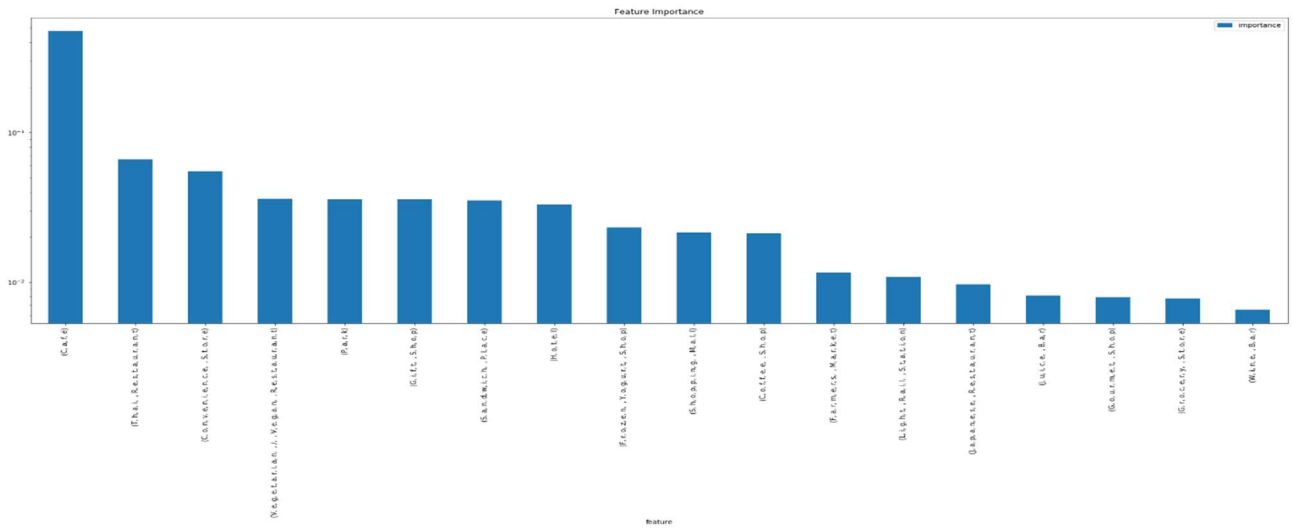


Random Forest Regression for type = town in N M Region

Random Forest Regression was carried out to determine relationship between venue category and average property price. In this regard data frame joined_venues_price_N_M_t_df was utilized which resulted in regression coefficient (R^2) of ~ 0.58 with mean squared error (mse) of ~ 2.8E+10. Grid Search CV was utilized to determine best parameters including max_depth, min_samples_leaf & n_estimators in order to optimize R^2 value. Below plot shows list price versus predicted price;



In addition to this, feature importance using random forest regression was performed to determine venues with respect to their importance on average property prices of neighbourhoods. Below is the plot showing venue with its importance with respect to average property price;



Results

This results section provides an overview of the outcomes of the methodology and their relevance to the original problem of identifying price variation with respect to housing attributes for different neighborhoods in Southern & Northern Metropolitan regions of Melbourne in conjunction with nearby venues to resolve the encounters faced by any individual during selection of a place.

- Histogram for all property type was generated in order to provide an insight of the pricing frequency in the city of Melbourne in general which lies between $ 800,000 - 120,000,0.
- Box plot was generated to identify the relationship between price & categorical variables but there seems an overlap between all variables, hence these variables can't be good predictors of price.
- Heat map & Pair plot suggest the strong to moderate relationship of price with number of rooms, distance from CBD, number of bathrooms & year built.
- Line plot between mean price of all three types of properties in Melbourne suggests that buyers should be interested in buying House in coming years due to dramatic change (~ $100,000/year) in price.
- Line plot between mean price of all 3 types of properties in Southern Metropolitan region suggests that sellers should be interested in selling Houses in coming years due to dramatic change (increased by ~ $100,000/year) in price.
- Line plot between mean price of all 3 types of properties in Northern Metropolitan region suggests that sellers should be interested in selling Townhouses in coming years due to dramatic change in price.
- GBR gives maximum R^2 of 0.62, 0.59 & 0.45 when applied on all variables for all type of properties, for type = house & for type = townhouse in Melbourne, respectively.
- Ridge regression gives maximum R^2 of 0.53 when applied on all variables for type = unit in Melbourne.
- GBR gives maximum R^2 of 0.75, 0.61, 0.65 & 0.53 when applied on all variables for all type of properties, for type = house, for type = unit & for type = townhouse in Southern Metropolitan region.
- GBR gives maximum R^2 of 0.72, 0.7 & 0.77 when applied on all variables for all type of properties, for type = house & for type = townhouse in Northern Metropolitan region.

- Maximum number of venues & unique venues category are concentrated in Northern Metropolitan region, where house & units are located.

	Regression type	R^2	MAE	MSE	RMSE	Yhat	
Melbourne	Linear	0.23	407809	3.474E+11	589435	724427	
		0.05	456927	4.264E+11	653013	882389	
		0.22	425552	3.525E+11	593727	1238825	
		0	479036	4.507E+11	671370	1081777	
	MLR	0.5	316917	2.237E+11	472982	482900	
	Polynomial Pipe	0.57	290103	1.949E+11	441452	444520	
	Ridge	0.55	284587	1.859E+11	431125		
	Lasso	0.52	308633	2.135E+11	462028		
	GBR	0.62	270466	1.587E+11	398392		
Southern Metropolitan Region	Linear	0.49	438260	4.058E+11	637025	886874	
		0	669042	7.945E+11	891349	1365098	
		0.39	517633	4.871E+11	697946	1651077	
		0.02	662206	7.787E+11	882458	1961055	
	MLR	0.63	365962	2.924E+11	540745	522907	
	Polynomial Pipe	0.68	338543	2.509E+11	500889	627889	
	Ridge	0.69	353303	2.807E+11	529781		
	Lasso	0.63	362870	2.893E+11	537895		
	GBR	0.75	314712	2.317E+11	481312		
All types in Northern Metropolitan Region	Linear	0.13	307566	1.71E+11	413471	725713	
		0.16	280010	1.64E+11	405030	570957	
		0.07	318820	1.825E+11	427244	975305	
		0	321918	1.957E+11	442327	866567	
	MLR	0.61	191768	7.717E+10	277798	331443	
	Polynomial Pipe	0.69	167021	6E+10	244950	492378	
	Ridge	0.7	151614	4.511E+10	212391		
	Lasso	0.61	195388	7.266E+10	269551		
	PCA	0.72	141530	4.181E+10	204465		

House type in all regions Melbourne

Regression type	R^2	MAE	MSE	RMSE	Yhat
Linear	0.12	482321	4.4435E+11	666598	829940
	0.16	457295	4.2457E+11	651592	929938
	0.17	468515	4.1831E+11	646768	1330695
	0.01	504354	4.9846E+11	706018	990582
MLR	0.45	359884	2.7484E+11	524248	517153
Polynomial Pipe	0.53	327849	2.3631E+11	486122	350900
Ridge	0.54	323986	2.4289E+11	492841	
Lasso	0.46	358092	2.8476E+11	533628	
GBR	0.59	308234	2.19E+11	467972	
Linear	0.26	513922	5.376E+11	733212	1197267
	0.06	589097	6.7955E+11	824346	1274496
	0.3	505210	5.1025E+11	714320	1930511
	0.01	610849	7.1993E+11	848487	2170160
MLR	0.48	422645	3.7598E+11	613172	532558
Polynomial Pipe	0.61	371753	2.858E+11	534598	870663
Ridge	0.61	388576	3.0172E+11	549292	
Lasso	0.48	386729	2.6819E+11	517867	
GBR	0.61	376565	3.0421E+11	551548	
Linear	0.02	337502	2.0372E+11	451358	884379
	0.39	239355	1.27E+11	356370	557449
	0.03	334647	2.0248E+11	449980	1036375
	0.02	329994	2.0491E+11	452669	795834
MLR	0.57	196862	8.5058E+10	291646	382061
Polynomial Pipe	0.7	167918	6.3257E+10	251509	400445
Ridge	0.65	182592	7.7665E+10	278683	
Lasso	0.61	195348	7.4304E+10	272588	
GBR	0.7	167309	6.5795E+10	256504	

Unit type in all regions in Melbourne

Regression type	R^2	MAE	MSE	RMSE	Yhat
Linear	0.27	139612	4.43E+10	210393	601713
	0	173580	6E+10	245034	558879
	0.22	154869	4.68E+10	216363	830029
	0	172374	6E+10	244951	669846
MLR	0.42	128069	3.5E+10	187039	666461
Polynomial Pipe	0.48	123458	3.13E+10	177005	616668
Ridge	0.53	123549	2.71E+10	164664	
Lasso	0.39	134973	3.66E+10	191342	
GBR	0.49	126081	3.05E+10	174704	
Linear	0.36	141024	4.74E+10	217751	676644
	0	190323	7.45E+10	273035	661398
	0.25	167457	5.6E+10	236562	924835
	0.01	189352	7.41E+10	272278	756801
MLR	0.49	129300	3.8E+10	194997	770772
Polynomial Pipe	0.53	123672	3.48E+10	186467	925145
Ridge	0.64	128885	2.95E+10	171897	
Lasso	0.38	140583	4.92E+10	221771	
GBR	0.65	126782	2.86E+10	169070	
Linear	0.22	124659	3.52E+10	187724	559261
	0.08	139802	4.13E+10	203115	330682
	0.24	126641	3.41E+10	184537	774839
	0.01	144254	4.47E+10	211448	611403
MLR	0.51	100431	2.21E+10	148707	520936
Polynomial Pipe	0.63	90004	1.68E+10	129569	425076
Ridge	0.6	91011	1.95E+10	139656	
Lasso	0.53	94326	1.88E+10	137100	
GBR	0.56	92110	2.16E+10	146907	

Unit type in SM Region

Unit type in NM Region

Townhouse type in all regions in Melbourne

Regression type	R^2	MAE	MSE	RMSE	Yhat	
Linear	0.19	244730	1.219E+11	349186	701027	2 rooms
	0.02	279363	1.475E+11	384038	808019	20 km distance
	0.2	245754	1.201E+11	346522	968423	2 bathroom
	0.01	279083	1.495E+11	386711	1029662	year 2019
MLR	0.34	221245	9.857E+10	313953	641096	
Polynomial Pipe	0.4	209798	8.963E+10	299386	558819	
Ridge	0.39	223910	9.587E+10	309634		
Lasso	0.35	234859	1.019E+11	319204		
GBR	0.45	210631	8.81E+10	296813		
Linear	0.18	283960	1.613E+11	401599	901739	
	0.06	300037	1.852E+11	430300	925264	
	0.16	287498	1.641E+11	405077	1200950	
	0	310288	1.921E+11	438272	1456648	
MLR	0.37	238117	1.229E+11	350571	672380	
Polynomial Pipe	0.48	217177	1.027E+11	320546	681997	
Ridge	0.44	248795	1.287E+11	358786		
Lasso	0.35	268272	1.507E+11	388232		
GBR	0.53	239904	1.083E+11	329064		
Linear	0.04	192099	6.234E+10	249679	703287	
	0.49	132854	3.301E+10	181679	294654	
	0.12	183837	5.669E+10	238089	811421	
	0.01	196135	6.427E+10	253522	819872	
MLR	0.66	109784	2.177E+10	147552	411856	
Polynomial Pipe	0.75	94928	1.602E+10	126574	559303	
Ridge	0.61	117792	2.646E+10	162673		
Lasso	0.58	115083	1.953E+10	139734		
GBR	0.77	89951	1.253E+10	111941		

Townhouse type in S M Region

Townhouse type in N M Region

House Type Southern Metropolitan Region

House Type Northern Metropolitan Region

Discussion & Recommendation

As a result of subject analysis, following trail can be followed in order to find suitable place to live;

- For any newcomer, the main area of interest for finding accommodation should be the neighborhood located in Northern Metropolitan region.
- Secondly, property type condo should be the focus due to its affordability in comparison to other property types.
- Thirdly, number of bathrooms requirement to be optimized as per family need as it has a moderate to strong association with property price.
- Lastly, consider those neighborhoods which are without or at least have minimum number of venues such as cafes, dessert shops & fast food restaurants as they have a major impact on property price.

One more key fact which has always kept Melbourne's real estate market in a bubble is due to less supply of accommodations in comparison to demand. This is due to continuous influx of locals from other provinces of country and immigrants from all around the globe which has led increase in Melbourne population with a smaller number of places to live. Since Australia aims to increase invitations for more immigrants in coming years from all around the globe which will keep rising Melbourne population leaving real estate market hype.

Conclusion

Since the objective of the project was to investigate the relationship between property attributes and nearby venues with respect to property price. As a result of subject study and analysis done with integration of real estate and location data, following are the outcomes and observations;

- Exploratory data analysis in conjunction with modelling suggests bathroom count as key attribute for property (house & unit) price prediction in Southern Metropolitan.

- Based on data visualization, neighborhoods closer to the Southern Metropolitan region in Melbourne are the expensive ones possibly due to less venue types.
- Based on data visualization, neighborhoods closer to the Northern Metropolitan region in Melbourne are the least expensive ones possibly due to greater number of venue types.
- With respect to number of venue types and category, price for townhouse & house in Northern Metropolitan region shows reasonably good relationship with R^2 value of 0.58 & 0.4, respectively.
- With respect to modelling including Multilinear, Polynomial, Lasso and Ridge regression shows similar results in terms of regression coefficient and mean squared error, with exceptions of Unit & Townhouse types in Southern Metropolitan region.
- Since regression techniques including linear and polynomial regression didn't show any association of venue category data with neighborhood average property price. In this regard, random forest regression with feature importance was performed which improved regression coefficient and showed feature importance with respect to average property price of neighborhoods.
- Based on feature importance results; cafes, Australian & Japanese Restaurants are considered as top three venues that have major impact on average property price of townhouse in Southern Metropolitan region.
- Based on feature importance results; cafes, convenience store & pizza place are considered as top three venues that have major impact on average property price of unit in Southern Metropolitan region.
- Based on feature importance results; cafes, light rail station & home service place are considered as top three venues that have major impact on average property price of house in Southern Metropolitan region.
- Based on feature importance results; cafes, Thai restaurant & convenient store place are considered as top three venues that have major impact on average property price of townhouse in Northern Metropolitan region.
- Based on feature importance results; Japanese restaurant, Sculpture Garden & Cafe are considered as top three venues that have major impact on average property price of unit in Northern Metropolitan region.
- Based on feature importance results; Wine bar, Café & Dog run are considered as top three venues that have major impact on average property price of house in Northern Metropolitan region.

So, in a nutshell, locals moving to Melbourne from other provinces of Australia and international immigrants planning to settle in Melbourne should consider above mentioned factors as qualifiers for selection of an appropriate place for living.