Predicting Condo prices in Sao Paulo

Napattarapon Pranmontri

September 11, 2019

1. Introduction

Background

The real estate market is one of the most competitive in terms of pricing and the same tends to vary significantly based on a lot of factors, hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. Therefore in this project, we present various important features to use while predicting housing prices with good accuracy. We have described regression models, using various features to have lower Residual Sum of Squares error.

Inspiration from data provider

This is a smaller and anonymized version of a dataset they use on thier startup https://openimob.com which predicts the price of properties and help people to find the best deals on the real estate market.

Objective

Our main goal is to building a machine learning model for predicting condo price in Sao Paulo to help real state investors to make better business decisions.

- Build a simple regression model for predicting land prices in Sao Paulo.
- Improve prediction model by adding data through the Foursquare API.

2. Description of the data

The main data used for this project will be from two sources:

- The rent/sale condo price in Sao Paulo April 2019. (Kaggle)

  This dataset contains around 13.000 apartments for sale and for rent in the city of São Paulo, Brazil. The data comes from multiple sources, specially real estate classified websites.

- The venues in each neighborhood. (FourSquare API)

Link to kaggle Dataset

Limitations

- Due to limited calling API from Foursquare each times, we can't use all sale data from the original(6412rows) so in this project, we will focus on recent property marked as 'New' = 1 in the original dataset.

Columns Meaning

- PriceFinal price advertised (R$ Brazilian Real)
- CondoCondominium expenses (unknown values are marked as zero)
- SizeThe property size in Square Meters m² (private areas only)
- RoomsNumber of bedrooms
- ToiletsNumber of toilets (all toilets)
- SuitesNumber of bedrooms with a private bathroom (en suite)
- ParkingNumber of parking spots
- ElevatorBinary value: 1 if there is elevator in the building, 0 otherwise
- FurnishedBinary value: 1 if the property is funished, 0 otherwise

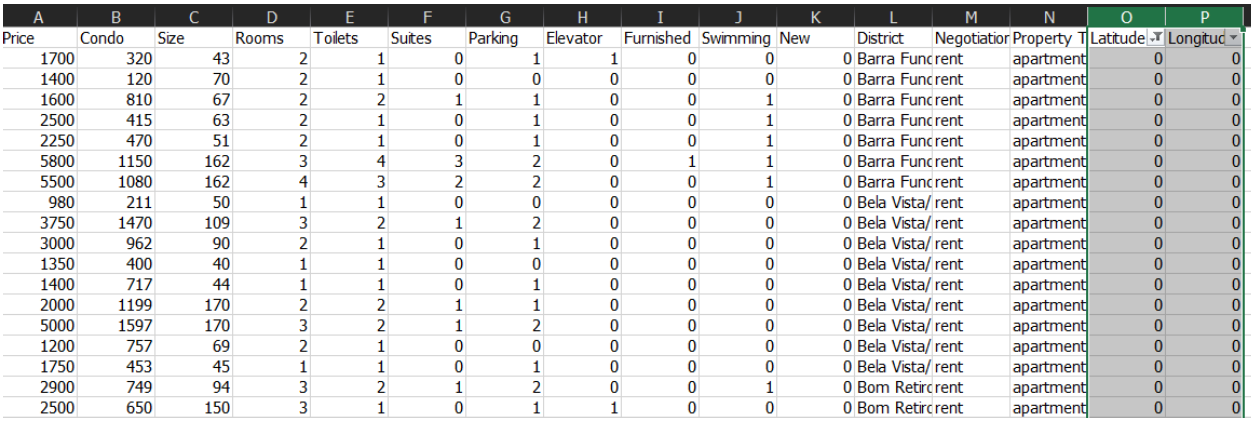
- Swimming PoolBinary value: 1 if the property has swimming pool, 0 otherwise
- NewBinary value: 1 if the property is very recent, 0 otherwise
- DistrictThe neighborhood and city where the property is located, e.i: Itaim Bibi/São Paulo
- Negotiation TypeSale or Rent
- Property TypeThe property type
- Latitude & Longitude

3. Methodology

Firstly,we explore the dataset and extract only useful information for further analysis(cleaning the data and create new feature from Foursquare API are also included in this part).Secondly, correlation between sale price and each factors will be checked.Second, if correlated, machine learning techniques (Multiple Regression) will be used to analyze the data.Finally,we compare the accuracy between each the model combinations for finding the best model with highest accuracy.

3.1 exploratory data analysis

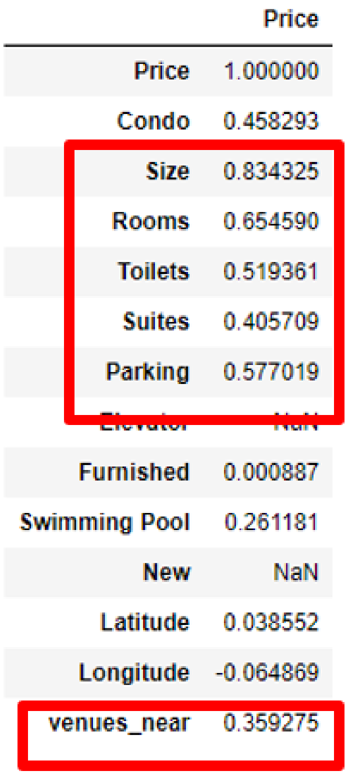
Clean the original data set



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Price	Condo	Size	Rooms	Toilets	Suites	Parking	Elevator	Furnished	Swimming	New	District	Negotiatior	Property T	Latitude	Longitud
1700	320	43	2	1	0	1	1	0	0	0	Barra Func	rent	apartment	0	0
1400	120	70	2	1	0	0	0	0	0	0	Barra Func	rent	apartment	0	0
1600	810	67	2	2	1	1	0	0	1	0	Barra Func	rent	apartment	0	0
2500	415	63	2	1	0	1	0	0	1	0	Barra Func	rent	apartment	0	0
2250	470	51	2	1	0	1	0	0	1	0	Barra Func	rent	apartment	0	0
5800	1150	162	3	4	3	2	0	1	1	0	Barra Func	rent	apartment	0	0
5500	1080	162	4	3	2	2	0	0	1	0	Barra Func	rent	apartment	0	0
980	211	50	1	1	0	0	0	0	0	0	Bela Vista/	rent	apartment	0	0
3750	1470	109	3	2	1	2	0	0	0	0	Bela Vista/	rent	apartment	0	0
3000	962	90	2	1	0	1	0	0	0	0	Bela Vista/	rent	apartment	0	0
1350	400	40	1	1	0	0	0	0	0	0	Bela Vista/	rent	apartment	0	0
1400	717	44	1	1	0	1	0	0	0	0	Bela Vista/	rent	apartment	0	0
2000	1199	170	2	2	1	1	0	0	0	0	Bela Vista/	rent	apartment	0	0
5000	1597	170	3	2	1	2	0	0	0	0	Bela Vista/	rent	apartment	0	0
1200	757	69	2	1	0	0	0	0	0	0	Bela Vista/	rent	apartment	0	0
1750	453	45	1	1	0	1	0	0	0	0	Bela Vista/	rent	apartment	0	0
2900	749	94	3	2	1	2	0	0	1	0	Bom Retirc	rent	apartment	0	0
2500	650	150	3	1	0	1	1	0	0	0	Bom Retirc	rent	apartment	0	0

We indicate that some data lack of location in formation which is required for extracting feature by using Foursquare API so we decide to not consider this data in this project.

Pearson correlation Table

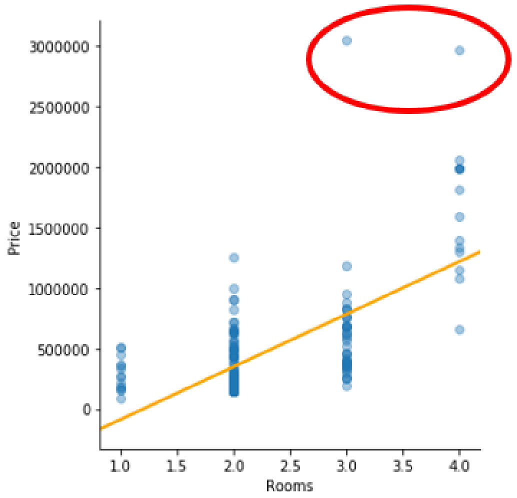


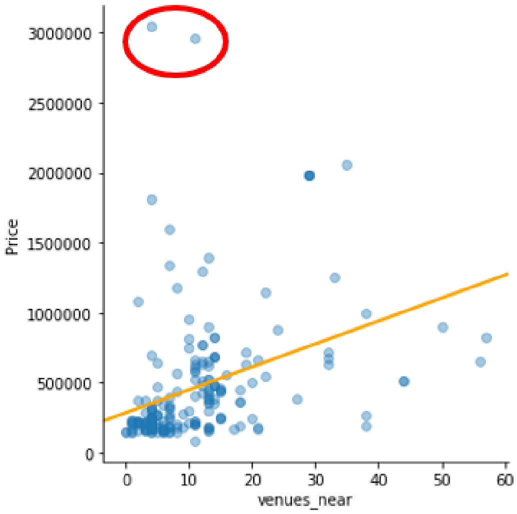
	Price
Price	1.000000
Condo	0.458293
Size	0.834325
Rooms	0.654590
Toilets	0.519361
Suites	0.405709
Parking	0.577019
Elevator	NaN
Furnished	0.000887
Swimming Pool	0.261181
New	NaN
Latitude	0.038552
Longitude	-0.064869
venues_near	0.359275

- We can clearly see that 'Size' related to 'Price' significantly.

- 'Rooms' 'Parking' and 'venues_near' seem slightly related to 'Price'.

Outlier detection





3.2 machine learnings

Preparation

We have used Multiple linear regression models and have calculated the root mean squared error and R square for each. Graphs have been plotted for each model. he data-set we have used is Sao Paulo Real Estate - Sale / Rent - April 2019 The size of the dataset is of 195 condos which are divided into training data and testing data in the ratio 80:20 .The number of features present in our data is square feet, price, a number of rooms, floors, a number of parking lots and a number of near venues.

Multivariate regression models:

In multivariate models instead of 1 feature, we use several features for better accuracy.

Model1 = Regression trained using [square feet,rooms, parking]

Model2 = Regression trained using [square feet,rooms, parking,near venue]

Next,we compare mean squared error ,root mean squared error and R square between each combinations. Finally after we have got the best model, we recommend any suggestion for model improvement.

4. Results

Model1

- R Square = 0.897
- MSE = 13,160,500,765.816427
- RMSE = 114,719.22578982316

Model2 (add ‘near venues’ feature)

- R Square = 0.932
- MSE = 8,703,312,256.41301
- RMSE = 93,291.54439933455

From results above, we can conclude that model2 with including ‘near venues’ feature is better than model1.

5. Recommendation

The following are suggestions how this project could be further developed:

- Due to Foursqure limitation in this project,we should use all the dataset(6412rows) from kaggle for building a model.

- Due to wealth difference of each districts in sao paulo,we should categorize the wealth level and then create a model based on each wealth level.

- Find the other features that may affect the condo price for better accuracy.For example,How old,facility rating from resident,etc.