Final Capstone - Predicting Real Estate Prices

Hector Hernandez

August 11, 2020

1 Introduction

1.1 Background

Many times we ask ourselves what the right price is for a real estate listing. To determine the ”just” or ”fair” price of real estate we usually rely on data regarding similar listings and build an idea from there. We tend to do this without much analytical rigor or without a measure of how proper our analysis is. It is in this issue that I find an opportunity to build a predictive model for real estate prices; therefore, I will be focusing on building a supervised machine learning model to predict the prices of real estate locations in Mexico City.

1.2 Stakeholders

The question the model seeks to answer is what is the ’fair price’ or ’expected price’ for a real estate location given a set of features or characteristics such as area, the number of rooms, etc. This model can serve a lot of purposes; it can help real estate agencies in the pricing of new listings, it can help customers get an idea as to the price of real estate they might be searching for, it can flag relatively cheap real estate, etc. In particular, to build the model the Foursquare API can be hugely beneficial because it can identify commercial locations around the real estate. The commercial locations around a real estate likely influence the price of it which is why with foursquare data we can build a better predictive model.

2 Data Sources

2.1 Real Estate Listings - Mexico City

To build the model mentioned above I will be using mainly two data-sets. The first data-set contains real estate listings from a real estate agency’s web-page. This data set was built by scraping the following url: `https://www.remax.com.mx/propiedades/ciudad+de+mexico_ciudad+de+mexico/venta`. This data set contain the features or characteristics of listed real estate properties such as number of rooms, area, neighborhood and so forth; it also contains the price of each listing. The process of scraping the data was long and tedious and besides the point in this capstone, so I converted the data to a CSV and loaded it into this repository.

2.2 Foursquare Venue Data

The second database I use, which is merged with the previous data set, is foursquare venue data which contains commercial venues around a defined radius of each listing in the previous data set. To get this data I will use the foursquare API and the proper queries. Because commercial venues around a property might affect its price it seems like a perfect opportunity to include more features in the data set.

2.3 Data Cleaning

The first step in data manipulation consisted in cleaning the addresses such that the geolocator package could find the corresponding latitude and longitude for each listing. The format of the address had to be properly set up otherwise the geolocator would return a null value. Unfortunately, the process of correcting addresses is too painstaking for the scope of this Capstone; out of the 250 original listings the geolocator only returned 168 —after considerable effort at writing the addresses properly—.

The second step consisted of making a foursquare API query to get all the commercial venues in a 500 meter radius of the listed locations latitude and longitude. This data set was grouped by address since there are multiple commercial venues per address; there were a multitude of different commercial venues, therefore including a dummy for each one may have led to an over-fitting model, I therefore aggregated the commercial venues without distinguishing them. This data set was merged with the other features of the listed real estates.

3 Methodology

I used a linear regression model with real estate listing price as the dependent variable and the listing's features as independent or predicting variables. I chose a linear regression model because price is a continuous variable. The followings sections describe the process to estimate the model.

3.1 Data Processing

Because it is fairly obvious that location is a substantial predictor of price it would seem like bad practice to exclude it. However, because there are so many distinct neighborhoods the model may be prone to over-fitting if we include all existing neighborhoods as dummies. For this reason I clustered neighborhoods based on the k-means clustering unsupervised algorithm. I clustered them based on their geographical location since it seems reasonable the nearby listings should be similarly priced (all else being equal). I tried several number of clusters —from two to nine clusters— and ended up choosing the number of clusters which provided the best performance in an out of sample data.

I tried standardizing the data but the power of the model didn't increase substantially so in the end I used the raw data. I also tried using polynomial features, but because of the number of observations, the model was over-fitting. The best model was therefore a simple linear regression with all the relevant features including: area, neighborhood cluster, number of rooms, number of baths, number of parking spaces, type of listing and the number of commercial locations around the property.

3.2 Goodness of Fit

Lastly, I split the data into a training (70%) and testing set and calculated the R^2 on the testing set as a measure of the goodness of fit.

4 Results

The best out of sample R^2 that I managed to get was equal to .34 with a 7 neighborhood cluster and the rest of the features. I used 30% of the data for testing and the rest for training the model. The coefficients for the model can be seen in the table below. The reference category for "type" is "Terrain" and the reference category for "N_cluster_i" is cluster 0.

Table 1: Coefficients in Real Estate Price Regression

room	0.0
bathroom	-841581.94
m2	331197.41
parking	19633.38
commercial_count	2974303.56
type_Bodega	-30713.33
type_Casa	-8791145.54
type_Comercial	-11438220.61
type_Departamento	-10165014.68
N_cluster_7_1	-6870873.17
N_cluster_7_2	-226706.43
N_cluster_7_3	-2145533.37
N_cluster_7_4	-3373129.76
N_cluster_7_5	-1672112.90
N_cluster_7_6	-2902659.77
Observations	168
Out of sample R^2	0.34

5 Discussion

At first the model was prone to over-fitting when I standardized the data and when I included polynomial features; this stands to reason because there are only 168 observations so it is very likely that as we included polynomials in the regression the model was prone to over-fitting. In the training data the R^2 was considerably bigger than in the testing data. I found that the simple linear model with 7 distinct neighborhoods —based on geographical proximity— and the rest of the features was the best out-of-sample predictive model with an R^2=0.34.

The best insight we got from the data is that neighborhood zero when making seven cluster is the most expensive neighborhood and that the number of commercial venues is a significant predictor of price; the more commercial venues there are the higher the price.

6 Conclusion

In this capstone I set about building a predictive model in order to predict the price of a real estate listing given its features. To build the data set I scraped a web-page for their real estate listings and then I used the foursquare API to get the nearby commercial venues. Next, I searched through different linear regression models, including polynomial features and different neighborhood clusters, and found many of them to over-fit the data; I ended up using a normal linear regression with no polynomial features and with 7 clusters of neighborhood as the best predictive model.