Recommender System Based on Millennial Preferences in the Real Estate Sector in Mexico's Capital

Alec García Barba

April 2020

1. Introduction

Background

Given that the 'millennial' generation is known, among many other aspects, to be a generation that prefers renting over buying, and that it is slowly moving towards an age where marrying, having kids and settling are the expected steps for this age range that varies from 1980s to 1996.

A study realized by the company *iCasas* [1] *showed that* there are approximately 30 million people in Mexico within the generation Y. And Mexico's capital is where 80% of young adults rent, and 20% of them chooses to buy its first home. The second in list is the State of Mexico, which is directly at north of CDMX, with almost five million young adults renting or buying. This region is by far the best suit to implement a solution to this market needs.

Problem Description

Selling houses and departments to this generation can be tough. This technology and freedom driven generation seems to be unreadable to the real estate sector, making the real estate business have a hard time renting and selling to this niche. Factors like a big garden, 3 bathrooms, or 3 story houses are not as relevant to potential customers as they used to be in other generations, like boomers, for example. A new approach to real estate's sales strategy must be reached.

Business Opportunity

I believe that data can generate a profile that fits a large portion of the financially stable person, between the ages 25 – 37 approximately, and generate a cluster of cities that matches its interests as well as benefit its lifestyle. With this clustering segmentation, a recommender system can be applied to the neighborhoods and finally deliver a set of neighborhoods which can be potentially profitable to the sector.

This project aims to find the best neighborhoods for this particular generation section profile, in order to seek business opportunities in the real state sector.

Previous analysis

A survey conducted by NAR[2] showed that 62% of millennials would rather have easy access to restaurants and shops, as well as shorter commutes than a bigger home. This survey also showed that an overwhelming 88% of the sample population believes that having amenities within walking distance increased quality of life. 70% also said that walkability, short commutes and proximity are important factors when deciding where to live. On the contrary, the majority of older generations consider that they have no problem with a longer commute and driving to amenities, if it means living in a single-family, detached home.

2. Data Acquisition and cleaning

Acquisition

At least two different datasets will be used. The first one will be a public dataset provided by government, containing all the geographical coordinates, neighborhoods and Boroughs of the city[3]. This data is vital to the project. The second dataset will be obtained via Foursquare, containing all the nearby venues of each neighborhood, obtained based on the coordinates of the first dataset. I believe these two sets will provide sufficient information about all the amenities desired by our profiled buyer, considering transportation, amusement (cultural and social), restaurants and a general sense of living costs near the neighborhood.

In case these two datasets were insufficient to compare and analyze each neighborhood, two additional datasets will be added to our base. One containing information regarding crime (by neighborhood and Borough)[4] , and one containing information regarding the position of every choice of public transportation.[5] This will be to obtain new insights in security, and a complete scheme of transportation options that would benefit potential customers.

Cleaning

The information of the first dataset will be filtered severely. Columns regarding the entity number, postal code, neighborhood ID, etc. will be dropped. The information of this dataset will only be the neighborhood, its Geolocation and geo shape, and the Borough.

Given the size of the capital, and the limitation on my Foursquare developer account, this set will also be filtered to the top 5 boroughs in CDMX, based on population, commuting and general popularity, this boroughs are: *Coyoacán, Benito Juárez, Miguel Hidalgo, Tlalpan and Álvaro Obregón.* I will obtain nearby venues via the Foursquare API and focus primarily on public transport, then on

social and cultural venues.

In case of merging security and complete transportation datasets, the security dataset would be grouped by neighbor and counted, in order to have a numeric value representing crime in each neighborhood. Lastly, the public transport system dataset would be merged by its coordinates, and using geo shape information, each station would be assigned to a neighborhood. This is a last resource in case the foursquare API could not deliver public transportation venues as accurate as desired.

References

1. ICASAS. Los 5 estados donde viven los millenials. Available at: https://www.icasas.mx/noticias/donde-viven-los-millenials/.
2. NAR. 2017 National Community and Transportation Preference survey. (2017). Available at: https://www.nar.realtor/reports/nar-community-and-transportation-preferences-surveys.
3. CDMX GOB. No TiBase de datos de las Colonias de la Ciudad de Méxicotle. (2020). Available at: https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/table/.
4. CDMX GOB. Carpetas investigacion CDMX. (2020). Available at: https://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-de-la-ciudad-de-mexico/.
5. CDMX GOB. Paradas y Terminales del Sistema de Transporte Publico. (2020). Available at: https://datos.cdmx.gob.mx/explore/dataset/estaciones-paradas-y-terminales-del-sistema-de-transporte-unificado/table/.