Clustering of Redfin Houses – your next perfect house

Shanshan Liang, Coursera Learner

Draft on 2018-10-29

# Abstract

# Introduction

Buying a house is a critical decision people make in various stages of lives. When people move into the new areas for jobs or family reasons, they settle down by embarking the house hunting journey. When life situation changes, they need a bigger house or a smaller house in the existing neighborhood that they’ve already fallen love with, they start the house hunting again. House hunting is important without doubt, yet a very complicated and resource intensive process. It is complicated due to people take various sets of factors in the decision making process – for example, someone may value the school district a lot, while as other people may not care much. There are many aspects of the houses people need to evaluate, with typically hundreds of houses on the market for research, all makes it a little intimidating in terms of the amount of the houses people need to go through before making a decision.

During this course, we’ve learned how to use FourSquare API to collect nearly venues data to cluster the neighborhoods. In this capstone project, I am leveraging this tool, together with the Redfin website to collect selected/custom information about houses for sale in my own zip code, as an example, to illustrate that house hunting process can be streamlined, customized and flexible. I hope this idea/process can help people gain more information about the targeted housing market, save time and increase the comfort level after fully analyzing the house market.

# Methods

Unsupervised machine learning model of clustering helps to group items together without the need of the dependent variable, or a label before running the algorithm. It has applications such as customer segmentation, identify groups to better understand the data. The feature set of the houses are important inputs in determining the cluster outputs. In this study, I purposely collect selected aspects of the houses rather than every available information from the Redfin website - based on what I value the most in buying a house. In real life situations, I expect the stakeholder would tell the data scientist what they value the most in decision making process. Similarly, only selected venue types are chosen to feed into the model based on the fact that most people are interested in a limited number of venues. In real life application, we could let the customer define or choose the venue types that they are most interested in.

Redfin data is collected via web scraping in Python. First, we identify the geography of interest, in my case, is my current 5 digit zip codes. In real life, it could be a new area someone will relocate to, it could be the zip code of the job site if short commute is important, it could be a broader area if someone wants to fully analyze the nearby areas. Second, we collect the information about each house in the search results. Web addresses of the houses for sale are saved in a list and we loop through this list to get information of each house. The features that I am particularly interested are total price, square feet, school rating and distance. Basic filtering can be applied such as I am only interested in houses that at least have 4 bedrooms and 2 baths. To use the FourSquare API, geo coordinates of the latitude and longitude of each house are also collected.

The surrounding venue data is collected via FourSqaure API using the explore query. Depending on the density of the interested venue, one could adjust the radius when query the place so it returns the desired number for the analysis. After the venue data is collected for each house, in the result set, I filter the venue type based on the understanding of my life pattern – the places that I go to most frequently already or interested to spend time on.

Redfin house data and the venue data is then merged to form the feature set of the houses for the clustering modeling. In order to determine the number of clusters, we will employ the “elbow method” and choose the number of cluster from the turning point.

# Results

# Discussion

# Conclusion