Real Estate Investment & California Price Prediction

Thip Rattanavilay

DSC680 - Fall 2021

https://github.com/thiprattanavilay/DSC680

Any surprises from your domain from these data?

There are plenty of surprise with the domain I have selected from the data gathered. I tried four different models, namely, linear regression, lasso regression, random forest and xgboost. The goal for this project is to predict the House Value of Houses in California's state and weather it is a smart place to invest into with the housing market today. The housing price predictions are usually conducted on a specific location and are difficult to generalize across different geographic regions.

The dataset is what you thought it was?

I knew that his was not going to be simple given the complex nature of this problem, to attain a model that accurately predicts the value of a home is difficult for a variety of reasons. One reason is that data sets with the complete information on all mentioned attributes are not easy to access. Secondly, the effect of a specific attribute on housing price may not have the same amount of significance on price variation across different regions. Pulling data from Zillow has paused because they have over bought homes (during 2021) and they don't want to share their data right now.

Have you had to adjust your approach or research questions?

Because Zillow has over bought homes, I will need to pull dataset from Kaggle "Zillow" from 2016-2018. I will try my best to present my findings to reflect the current events. Within the limited data available, my goals in this study are to assess 2 the prediction accuracy from various models and give clues of how certain features affect home prices. This information could prove useful to investors, buyers, and sellers of homes.

I will keep my research questions that I currently have made.

Is your method working?

I feel good about my method and of course in order for my method to work I will need find the right method to work with. I plan to stick with my current method and go from there.

Implementation approach:

1. Gather dataset
2. Handle missing values
3. Encode categorical data
4. Split the data
5. Standardize data

What challenges are you having?

Despite the discussed challenges, past research has shown that it is possible, at least partially, to assess the future value of a house. The nature of real estate markets is inherently intricate. In addition, the diversity of housing attributes makes it challenging to construct a comprehensive model that can encompass all of the features. The value of a house is substantially affected by its own unique set of structural attributes (e.g. living area, number of rooms, age of the building, swimming pool, garage), locational attributes (proximity to business districts and accessibility to major highways) and neighborhood attributes (e.g. school district, income levels, unemployment rate, population density).