Real estate regression study

Machine Learning - MSc DMDS - EM Lyon

Year 2020-2021

1 Introduction

Too often in real estate, the process of valuation can come across as a high-brow exercise of thumb-sucking. The realtor will come over and produce an estimated value with very little "quantitative" insight. Perhaps the process is exacerbated by the emotional attachment that owning property brings given that for many, a house will be the largest financial investment made in a lifetime.

Yet, there is a method to this madness, and this is what will be shown with this machine learning project. Using regression methods, we will show than we can make accurate predictions of the price of houses or terrain on new areas (that were not in the initial dataset) by using information about these areas.

2 Objectives / Tasks

The objective of the project is to determine the price of any house, apartment or terrain in any French city. For this study, you are proposed to use this dataset:

https://limmo-dvf.com/prix-immobilier

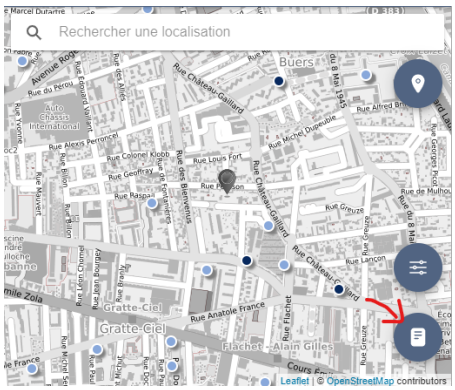
Which you can explore and which is an application that allows you to know the price of real estate sold in France. Navigate via the interactive map to find the properties in the area of your search. Limmo DVF is available as Web Application and Android Application.

The first task of this project is to determine which criterion could be impactful to decide the price of a certain house, apartment or terrain. Here are a few examples:

- Population of the city
- Distance to water (sea/lake)
- Public transportation (On a scale of your own)
- Population density
- Avg. Temperature

Be creative ! You should select at least 10 criterion. These choices should be made carefully, as these criterion should be: numerical, obtainable for any city (with any city size), decisive for the property price. You can later dismiss one of the selected criterion if you prove that it is senseless regarding your results. Don't forget to show this step during the final presentation.

The second task of this project is to determine a list of at least 10 areas that you will use to build your training dataset and to gather the information. From the limmo-dvf website, you can export the data of particular areas into excel. These information are



quite standard, and you will have to add information about the selected areas manually. I recommend using Pandas for this step, but you can use whatever tool for this.

The third task will be to train a machine learning model, which would analyze the dataset and predict, according to information of the new areas/cities, the price of new incoming offers of houses, terrains and apartments that would not be from your training dataset cities. You will have to prove the validity of your model.

The fourth task is to prepare a presentation of your results. This implies making visuals to explain your results. You will have 15 to 20 minutes to show the result of your work.

3 Grading

The grade of your project will be determined by the teacher after your presentation. There will be no individual grade for members of the same team, unless there are clear disparity in the effort that was put. The four different task will each represent 25% of the final grade.