Machine Learning For a House pricing Prediction Web Application

Alex Soudant

Ynov Ingésup M1,

20 Boulevard Général de Gaulle, 44200 Nantes

Correspondence: Alex Soudant. E-mail: alex.soudant@ynov.com

Abstract

In this proposal, I present a project for a web application that can generate predictions for housing prices in France. I wish to use Computer vision tools to help identify a pricing range mainly based on pictures from properties on sale. Examples of computer vision tools to test includes : VLFeat, Harris corner detector, SIFT, SURF, Fast corner detection, Kanade–Lucas–Tomasi feature trackers that would certainly be available in the openCV library in python programming langage. This approach permits an easy and fast way for potential house buyers/sellers to estimate a fair price for properties before putting it on sale. It is highly probable that a picture only estimation will be less accurate than adding complementary pieces of information such as the geographic localisation of the properties, its presence either in a city or in contryside, the property size and land around it. To select the model to estimate the property price, I will test a range of machine learning algorithms such as : regression methods, Support Vector Regression (SVR), k-Nearest Neighbours (kNN), and Regression Tree/Random Forest Regression. To train the selected model, I will need to put together a dataset of property pictures with a known price in France. I then could use it to fit local models with a variable geographic scope to test the importance of localisation in housing prices. As the training will occurs before deploying the web application and because most of the computation needed to estimate the pricing will happen on the server-side, I should be able to provide on the client-side a relative quick pricing estimation when a new property needs to be evaluated on the web application. Finally, I will compare our approach to existing housing price prediction models and select common criterions to show my progress in terms of prediction strength.
Key words : Machine learning, Computer vision, Housing price, prediction

Introduction

Housing is an important market for business purposes. Nowadays, potential buyers/sellers for properties have to manually visit a high number of housing agents websites to be able to estimate a range of prices. This is a complex and time consuming process and the information obtained may not be accurate. It is also known that the difference in prices between websites for the same property can be quite high and may cause a lack of understanding of the real market prices. Hence I propose to develop a tool to assist individuals into finding a more understandable price of property at locations of their interest in France.

To help determine these prices, I want to test if computer vision algorithms can help in the analysis process. Computer vision represents a chain of tasks to acquire, process, analyze and understand digital images, which will result with the extraction of high-dimensional data in order to lead into decision making.

Materials and Methods

The first step in my data mining procedure is to constitute a training sample from which we can gain insight of the visual attributes an image of a property on sale can provide to estimate its price. To our knowledege, at the time of proposal writing, there was no public dataset available already published to fulfill such a study on the determination of housing price based on property prictures. This is due to the fact that sell prices are private in France so we cannot use recent archives of estate pricing. Therefore, we hope to extract images from web scraping techniques like for instance, scrapy, an open source crawler base on python language. The ideal final dataset would be constituted of at least a few thousand images from known pricing and geographical position in France. Additionnal information such as property size and number of bedrooms could help improve accuracy of the modelling predictions but are optional to the initial project.

Usually, the data obtained from websites are not directly usable since real world data will have a lot of noise. To clean the dataset, we will need to remove invalid and empty entries then examine if duplicates exists. After data cleaning, we would like to use the tensorflow librairy in python to perform data analysis. Tensorflow offers a collection of machine-learning algorithms used in deep learning to perform deep neural networks research. Other classification algorithms such as random forests, Nearest Neighbour and Support Vector Regression (SVR) can be extracted from the sklean library.