Machine Learning For House Pricing Prediction in Loire Atlantique Region

ALEX SOUDANT

Ynov Ingésup M1,

20 Boulevard Général de Gaulle, 44200 Nantes

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

Abstract

I present here a project to generate predictions for housing prices in Loire Atlantique Region (France) based on pictures analysis from sale adverts. I used the openCV librairy for computer vision to prepare pictures followed by the use of Tensorflow library for deep learning analysis. The goal is to produce an easy and fast way for potential house buyers/sellers to estimate a fair price for properties before contacting real estate agents. Results show that predictions based solely on pictures are not accurate enough to predict prices of properties on sale. I tested different approaches to improve prediction accuracy that will be detailed further in this report. The best set up so far was to build a neural network with both pictures and house features (i.e. number of rooms, house surface, land surface and number of pictures) as parallel neurones on the same layer then merge the resulting tensors in an output layer. To gather pictures and features necessary to this analysis, a scrapper was built using Ember framework and ran over a house sale agency website. About 600 adverts were obtained for Nantes city which represent over 3000 pictures. Angers and Brest city were also scraped but not used in training or validation for the model.

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 price. 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 machine learning tool to assist individuals into finding a more understandable price from real estate.

Data Obtention

To our knowledge, 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, I built a javscript scraper using Ember framework to help building a decent dataset to perform analysis.

Data Preparation and Cleaning

After running the scraper to gather pictures from the website into different advert folders in the hard drive, I noticed that some pictures were corrupted. The main consequence in such pictures was that the resolution was pretty poor and uneven thoughout the picture. Therefore, it can be expected that these pictures contain less information than a correctely downloaded picture which can greatly affect the analysis when training the model to predict prices. A easy way to get rid of these images is to read the size of the picture by using opencv library and deleting every picture below a threshold of 9000 octets. Features of the houses on sale, extracted from the online adverts, were saved into one json file for each advert. Therefore, the first quality check was to assert if all advert folders had a json file and delete folders that did not contained one. Once, I was sure to have in each advert folder, a set of pictures and a json file, the features for every advert were regrouped into a csv file in a table format. Features and pictures were then paired into pickle files following tensorflow tutorial recommandations. After having each picture paired with features, I then checked for prices and house surface equal to zero and removed these data lines before analysis. Adverts with no price indicated seem to match properties that need to be directly negociated by contacting the real estate agency. When no house surface is indicated, a rapid check on the real estate website shows that some adverts concern land properties with no built house on it. Due to the fact that the scraper gather adverts from page 1 of the website to page 30, I expected that adverts from the beginning of the dataset could be of better quality (i.e. attrative prices, more pictures or more attractive features). Therefore, all adverts were first shuffled before segregating them into training and validation sets. The training dataset represents around 1900 paired pictures-features

which leaves about 700 paired pictures-features for validation. As I consider these numbers to be still quite low to train deep learning algorithms, no test dataset was built to keep more data for training and validation. Finally, all features were normalized to match a scale between -1 and 1. This is an essential transformation to allow the chosen optimizer to reach a minimum loss when running a graph session.

Data Obtention